MCDI
MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO

GERENCIA DE CAPITAL HUMANO

POSGRADOS

Especialista Virtual PLD

Propuesta de Intervención
Que para obtener el grado de
MAESTRO EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

Raúl Rufino Díaz

Asesor:

Dr. Sabino Miranda Jiménez

Dra. Obdulia Pichardo Lagunas

Ciudad de México, Octubre, 2023

CONACYT

INFOTEC POSGRADOS

Autorización de impresión

GOBIERNO DE MÉXICO | CONAHCYT CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS | INFOTEC

BIBLIOTECA INFOTEC
VISTO BUENO DE TRABAJO TERMINAL

Maestría en Ciencia de Datos e Información
(MCDI)

Ciudad de México, a 12 de enero de 2024

UNIDAD DE POSGRADOS
PRESENTE

Por medio de la presente se hace constar que el trabajo de titulación:

"Especialista Virtual PLD"

Desarrollado por el alumno: **Raúl Rufino Díaz**, bajo la asesoría del **Dr. Sabino Miranda Jiménez y la Dra. Obdulia Pichardo Lagunas** cumple con el formato de Biblioteca, así mismo, se ha verificado la correcta citación para la prevención del plagio; por lo cual, se expide la presente autorización para entrega en digital del proyecto terminal al que se ha hecho mención. Se hace constar que el alumno no adeuda materiales de la biblioteca de INFOTEC.

No omito mencionar, que se deberá anexar la presente autorización al inicio de la versión digital del trabajo referido, con el fin de amparar la misma.

Sin más por el momento, aprovecho la ocasión para enviar un cordial saludo.

Mtro. Carlos Josué Lavandeira Portillo
Director Adjunto de Innovación y Conocimiento

CJLP/jah

C.c.p. Felipe Alfonso Delgado Castillo.- Gerente de Capital Humano.- Para su conocimiento.
Raúl Rufino Díaz.- Alumno de la Maestría en Ciencia de Datos e Información.- Para su conocimiento.

Agradecimientos

A mi querida madre, Isabel, quien ha sido la razón y la fuente de amor que me da fortaleza a mi vida. Tus sacrificios, sabiduría, cariño, amor y lealtad siempre han sido mi guía constante.

A mis hermanos, Estela, Irais, Job y Rigoberto, quienes han compartido conmigo risas y momentos invaluables a lo largo de toda mi vida. Su presencia y confianza en mí me han impulsado a superar cualquier desafío.

A Adriana, quien estuvo a mi lado en cada etapa de este viaje académico y personal. Tu paciencia y comprensión me dieron la fuerza para seguir adelante.

A todos ustedes, gracias por creer en mí.

Tabla de contenido

Índice de figuras

Índice de cuadros

Siglas y abreviaturas

EDA	Análisis exploratorio de datos
CFT	Contra Financiamiento al Terrorismo
CHC	Cheques de Caja
CNBV	Comisión Nacional Bancaria y de Valores
CSNU	Consejo de Seguridad de las Naciones Unidas
DOF	Diario Oficial de la Federación
DCG	Disposiciones de Carácter General
FT	Financiamiento al Terrorismo
FDE	Formato de Dólares en Efectivo
FTI	Formato de Transferencias Internacionales
GAFI	Grupo de Acción Financiera Internacional
GAFILAT	Grupo de Acción Financiera de Latinoamérica
IF	Instituciones Financieras
LFPIORPI	Ley Federal para la Prevención e Identificación de Operaciones con Recursos de Procedencia Ilícita
ROIP	Reporte de Operaciones Internas Preocupantes
ROI	Reporte de Operaciones Inusuales
ROR	Reporte de Operaciones Relevantes
NER	Reconocimiento de Entidades Nombradas
PLD	Prevención de Lavado de Dinero
PLN	Procesamiento del Lenguaje Natural
SBR	Sistema de Búsqueda de Respuestas
SHCP	Secretaría de Hacienda y Crédito Público
SO	Sujetos Obligados
UIF	Unidad de Inteligencia Financiera

Glosario

Actividad Vulnerable: Se refiere a los actos, operaciones y servicios y pueden ser definidas como aquellas que por su naturaleza y características son susceptibles de ser utilizadas por sus clientes o usuarios para llevar a cabo actos u operaciones con recursos de procedencia ilícita [41].

Activos Virtuales: A la representación de valor registrada electrónicamente y utilizada entre el público como medio de pago para todo tipo de actos jurídicos y cuya transferencia únicamente pueda llevarse a cabo a través de medios electrónicos. En ningún caso se entenderá como activo virtual la moneda de curso legal en territorio nacional, las divisas, ni cualquier otro activo denominado en moneda de curso legal o en divisas [20].

Banca Múltiple: Institución financiera de intermediación que recibe fondos en forma de depósito de las personas que poseen excedentes de liquidez, utilizándolos posteriormente para operaciones de préstamo a personas con necesidades de financiación, o para inversiones propias. Presta también servicios de todo tipo relacionados con cualquier actividad realizada en el marco de actuación de un sistema financiero [14].

Cliente: A cualquier persona física, moral o fideicomiso que, directamente o por conducto de algún comisionista contratado por la Entidad respectiva actúe a nombre propio o a través de mandatos o comisiones, que sea cuentahabiente de una Entidad, o utilice, al amparo de un contrato, los servicios prestados por la Entidad o realice operaciones con esta [19].

Disposiciones de Carácter General: Requisitos que deben reunir los requerimientos de información y documentación que las autoridades judiciales, hacendarias federales y administrativas, a que se refieren los artículos 142 de la Ley de Instituciones de Crédito, 34 de la Ley de Ahorro y Crédito Popular, 44 de la Ley de Uniones de Crédito, 69 de la Ley para Regular las Actividades de las Socieda-

des Cooperativas de Ahorro y Préstamo, 55 de la Ley de Fondos de Inversión y 73 de la Ley para Regular las Instituciones de Tecnología Financiera, formulen a la Comisión Nacional Bancaria y de Valores [13].

Financiamiento al Terrorismo: Consiste en la aportación, financiación o recaudación de recursos o fondos económicos que tengan como fin provocar alarma, temor o terror en la población o en un grupo o sector de ella, para atentar contra la seguridad nacional o presionar a la autoridad para que tome una determinación [15].

Lavado de Dinero: Es el proceso a través del cual es encubierto el origen de los fondos generados mediante el ejercicio de algunas actividades ilegales o criminales (tráfico de drogas o estupefacientes, contrabando de armas, corrupción, fraude, prostitución, extorsión, piratería y últimamente terrorismo). El objetivo de la operación, que generalmente se realiza en varios niveles, consiste en hacer que los fondos o activos obtenidos a través de actividades ilícitas aparezcan como el fruto de actividades legítimas y circulen sin problema en el sistema financiero [16].

Oficial de Cumplimiento: Es un profesional encargado de asegurar que una empresa u organización cumpla con las leyes, regulaciones y normativas aplicables a su industria o sector y deberá ser un funcionario que ocupe un cargo dentro de las tres jerarquías inmediatas inferiores a la del director general de la Entidad [19].

Recomendaciones del GAFI: Son los estándares internacionales más reconocidos para combatir el Lavado de Dinero y el Financiamiento del Terrorismo (LD/FT) Las mismas incluyen una serie de medidas financieras, legales y de conducta que los países deben llevar adelante, en su mayoría basadas en instrumentos legales internacionales (convenciones de la ONU y de organismos supervisores) [22].

Reporte de Operaciones en Efectivo con Dólares: Reporte por cada operación de compra, recepción de depósitos, recepción del pago de créditos o servicios, o transferencias o situación de fondos, en efectivo que se realicen con dólares de

los Estados Unidos de América [19].

Reporte de Operaciones con Cheques de Caja: Reporte por cada Operación de expedición o pago de cheques de caja, realizada con sus Clientes o Usuarios por un monto igual o superior al equivalente en moneda nacional a diez mil dólares de los Estados Unidos de América [19].

Reporte de Transferencias Internacionales de Fondos: Reporte por cada transferencia internacional de fondos que, en lo individual, haya recibido o enviado cualquiera de sus Clientes o Usuarios durante dicho mes, por un monto igual o superior a mil dólares de los Estados Unidos de América o su equivalente en la moneda extranjera en que se realice [19].

Reporte de Operaciones Inusuales: Reportes de operaciones, actividades, conductas o comportamientos de un Cliente que no concuerde con los antecedentes o actividad conocida por la Entidad o declarada a esta, o con el perfil transaccional inicial o habitual de dicho Cliente, en función al origen o destino de los recursos, así como al monto, frecuencia, tipo o naturaleza de la Operación de que se trate, sin que exista una justificación razonable para dicha Operación, actividad, conducta o comportamiento, o bien, aquella Operación, actividad, conducta o comportamiento que un Cliente o Usuario realice o pretenda realizar con la Entidad de que se trate en la que, por cualquier causa, esta considere que los recursos correspondientes pudieran ubicarse en alguno de los supuestos previstos en los artículos 139 Quáter o 400 Bis del Código Penal Federal [19].

Reporte de Operaciones Interna Preocupante: Reporte de operaciones, actividades, conductas o comportamientos de cualquiera de los directivos, funcionarios, apoderados y empleados de la Entidad de que se trate con independencia del régimen laboral bajo el que presten sus servicios, que, por sus características, pudiera contravenir, vulnerar o evadir la aplicación de lo dispuesto por la Ley o las presentes Disposiciones, o aquella que, por cualquier otra causa, resulte dubitativa para las Entidades por considerar que pudiese favorecer o no alertar sobre la actualización de los supuestos previstos en los artículos 139 Quáter o 400 Bis

del Código Penal Federal [19].

Reporte de Operaciones Relevantes: Reporte de operaciones que se realicen con los billetes y las monedas metálicas de curso legal en los Estados Unidos Mexicanos o en cualquier otro país, así como con cheques de viajero y monedas acuñadas en platino, oro y plata, por un monto igual o superior al equivalente en moneda nacional a siete mil quinientos dólares de los Estados Unidos de América [19].

Sujetos Obligados: A las Entidades y a las sociedades o personas sujetas a las obligaciones a que se refieren los artículos 124 de la Ley de Ahorro y Crédito Popular, 71 y 72 de la Ley para Regular las Actividades de las Sociedades Cooperativas de Ahorro y Préstamo, 108 Bis de la Ley de los Sistemas de Ahorro para el Retiro, 91 de la Ley de Fondos de Inversión, 212 y 226 Bis de la Ley del Mercado de Valores, 492 de la Ley de Instituciones de Seguros y de Fianzas, 60 de la Ley Orgánica de la Financiera Nacional de Desarrollo Agropecuario, Rural, Forestal y Pesquero, 129 de la Ley de Uniones de Crédito, y 95 y 95 Bis de la Ley General de Organizaciones y Actividades Auxiliares del Crédito, exceptuando a los centros cambiarios, y 58 de la Ley para Regular las Instituciones de Tecnología Financiera [19].

Usuario: A cualquier persona física, moral o Fideicomiso que, directamente o a través de algún comisionista contratado por la Entidad respectiva al amparo del artículo 46 Bis-1 de la Ley y demás disposiciones aplicables, realice Operaciones con la Entidad de que se trate, o utilice los servicios que le ofrezca dicha Entidad, sin tener una relación comercial permanente con esta [19].

Introducción

La Comisión Nacional Bancaria y de Valores (CNBV) define el Lavado de Dinero como "*El proceso a través del cual es encubierto el origen de los fondos generados mediante el ejercicio de algunas actividades ilegales (siendo las más comunes, tráfico de drogas o estupefacientes, contrabando de armas, corrupción, fraude, trata de personas, prostitución, extorsión, piratería, evasión fiscal y terrorismo)*"[1], en otras palabras, es dar legalidad a las transacciones ilícitas vinculadas en cada una de estas actividades, realizadas principalmente por empresas fachada. Por lo que, tanto Sujetos Obligados (SO) así como Instituciones Financieras (IF) revisan los perfiles transaccionales de sus respectivos clientes o usuarios, para poder monitorear operaciones sospechosas, mismas que deben ser reportadas a la Unidad de Inteligencia Financiera (UIF) a través de la CNBV.

Para hacer este monitoreo transaccional, las IF y SO requieren de personal con un perfil altamente especializado para realizar el respectivo análisis e identificación de riesgo que tiene cada uno de sus clientes, y así desarrollar un criterio analítico que les permita determinar cuáles sujetos necesitan una investigación más a fondo.

Este perfil especializado necesita conocer toda la documentación oficial y legislación aplicable, tanto nacional como internacional sobre el tema, entre los que se encuentran, los estándares internacionales y recomendaciones por el Grupo de Acción Financiera Internacional (GAFI) y Grupo Egmont; la Guía para la prevención y detección de operaciones con recursos de procedencia Ilícita, dictada por la CNBV; tipologías sobre la simulación de lavado de dinero y financiamiento al terrorismo, generadas por la UIF, por mencionar algunos. Dependiendo del nivel de conocimiento, los analistas pueden fungir como instancias de consultas en materia de Prevención de Lavado de Dinero y Contra el Financiamiento al Terrorismo (PLD/FT), dando orientación sobre esta regulación, para dar cumplimiento a cada uno de los procedimientos y llevar a cabo una correcta aplicación de la normatividad.

[1] Comisión Nacional Bancaria y de Valores. Lavado de Dinero. (s.f.). https://www.cnbv.gob.mx/CNBV/Documents/VSPP_Lavado%20de%20Dinero.pdf

Problemática

En el ámbito de la Prevención de Lavado de Dinero y el Financiamiento al Terrorismo (PLD/FT), el desarrollo de un criterio analítico sólido es esencial para los analistas. Esto implica tener conocimiento sobre guías, lineamientos, tipologías, leyes, reglamentos y estándares nacionales e internacionales emitidos por diversas autoridades y organismos. A medida que se busca una mayor especialización en este campo, la curva de aprendizaje para los analistas tiende a extenderse considerablemente. La adquisición de conocimientos especializados es un proceso continuo y constante. Sin embargo, en la actualidad, no existe una herramienta interactiva o un modelo especializado sobre la regulación en PLD/FT en México.

Contar con un modelo especializado que facilite la búsqueda de información en el ámbito de PLD/FT conforme a la regulación en México sería de gran utilidad, ya que permitiría acceder de manera inmediata a procedimientos específicos y datos especializados relacionados con la prevención e identificación de operaciones con recursos de procedencia ilícita y financiamiento al terrorismo.

Por lo tanto, el propósito fundamental de este trabajo es poner a disposición un modelo especializado entrenado con un conjunto de datos enfocados en PLD/FT, de manera que cualquier persona interesada en adentrarse en este campo o profesionales que ya prestan sus servicios puedan consultar acerca de esta temática de manera más eficiente.

Objetivos

Objetivo general

El objetivo principal de este trabajo es crear y poner a disposición un modelo que permita responder preguntas en materia de Prevención de Lavado de Dinero y Financiamiento al Terrorismo (PLD/FT). Este modelo se basa en BERT y realiza extracción de

texto en documentos oficiales a través de un ajuste fino para predecir respuestas a partir de un párrafo.

Objetivos específicos

1. Identificar todos los documentos oficiales nacionales emitidos por las diferentes autoridades del Sistema Financiero Mexicano, así como los documentos internacionales emitidos por el GAFI y el Grupo Egmont, además de la información relacionada con las diferentes Actividades Vulnerables (AV).
2. Generar un repositorio de documentos con todas las Disposiciones de Carácter General emitidas, guías, lineamientos, tipologías, leyes, reglamentos, estándares nacionales e internacionales.
3. Establecer la metodología para la creación del conjunto de datos PLD/FT.
4. Crear un conjunto de datos de preguntas y respuestas para el entrenamiento y evaluación de algoritmos que permitan responder a preguntas realizadas por el usuario, el formato deberá cumplir los estándares para el entrenamiento con las librerías BERT y SQuAD.
5. Validar preguntas del conjunto de datos en materia de Prevención de Lavado de Dinero y el Financiamiento al Terrorismo con analistas especializados dentro de esta regulación.
6. Análisis exploratorio de datos para la información contenida dentro de cada uno de los documentos del repositorio y el conjunto de datos.
7. Entrenar y evaluar el desempeño de modelos BERT que permita dar respuestas a las preguntas realizadas por el usuario.
8. Comparar los resultados obtenidos para seleccionar el modelo con mejor rendimiento.

Justificación

El principal objetivo de este trabajo es poner a disposición un modelo entrenado en el dominio de PLD/FT, especializado para la regulación en México, para cualquier persona interesada en el tema. El propósito de este modelo es facilitar la consulta de información en la documentación oficial, con el fin de reducir los tiempos de investigación. Para ello, el modelo se ha entrenado utilizando un conjunto de datos en esta materia, construido a partir de la documentación oficial utilizada para la certificación por parte de la CNBV y la UIF, de tal manera que el *Especialista Virtual PLD* sea un modelo de sistemas de búsqueda de respuestas entrenado con fuentes confiables en materia de PLD/FT.

Delimitación

Este proyecto está diseñado particularmente para inquirir en el tema de PLD/FT, por lo que se crea un modelo de sistema de búsqueda de respuestas de dominio cerrado o restringido, donde sus respuestas consisten en segmentos de textos ubicados en los documentos oficiales disponibles, para ello, realiza una extracción de texto(s) de cada uno de los archivos que se encuentren en el repositorio creado para el objetivo de este proyecto.

Por otra parte, sólo podrá dar respuesta en el idioma español y leerá documentos en formato PDF, de manera que, la información que se vaya incorporando y que no esté en dicho formato, debe ser creada en estos archivos para poder ser procesado por el sistema de manera exitosa.

Capítulo 1

Marco teórico

Capítulo 1.

Antecedentes

1.1. Marco teórico

A continuación, se presentan conceptos relevantes asociados al Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) y los Sistemas de Búsqueda de Respuestas (SBR) con técnicas de Procesamiento de Lenguaje Natural (PLN), que son base para el desarrollo de este proyecto.

1.1.1. Análisis exploratorio de datos

El análisis exploratorio de datos (EDA, por sus siglas en inglés) es utilizado para analizar e investigar conjuntos de datos y resumir sus principales características con el objetivo de identificar valores atípicos u observaciones inusuales, revelar patrones y comprender posibles relaciones entre variables [10].

1.1.1.1. Clustering

El análisis de grupos o *clustering* es una técnica de análisis de datos en la que se busca identificar grupos o clústeres de objetos que son similares entre sí y diferentes a los objetos de otros grupos [5]. K-Means es uno de los algoritmos de aprendizaje no supervisado más utilizados en el analisis de datos, este algoritmo agrupa en k grupos a partir de sus características y todos los elementos asignados a un mismo centroide forman un clúster. Uno de los componentes esenciales del *clustering* es la medida de distancia, que determina qué tan similares o diferentes son los puntos de datos en función de una métrica específica. Las medidas de distancia más habituales son:
Distancia Euclidiana:

$$d_{euclidiana}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Donde:

- x_i y y_i representan los elementos de los vectores x e y.
- $\sum_{i=1}^{n}(x_i - y_i)^2$ denota la suma de los cuadrados de las diferencias entre las elementos de los vectores x e y.
- $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ representa la raíz cuadrada de la suma de los cuadrados de las diferencias. Esto finalmente da la distancia euclidiana entre los puntos x e y.

Distancia Manhattan:

$$d_{manhattan}(x, y) = \sum_{i=1}^{n} |(x_i - y_i)|$$

Donde:

- x_i y y_i representan los elementos de los vectores x e y.
- $|(x_i - y_i)|$ representa el valor absoluto de la diferencia entre los elementos de los vectores x e y. Esto asegura que la distancia sea siempre no negativa.
- $\sum_{i=1}^{n} |(x_i - y_i)|$ denota la suma de los valores absolutos de las diferencias entre las elementos de los vectores x e y.

Distancia Similitud Coseno:

$$d_{coseno}(x, y) = 1 - \frac{|\sum_{i=1}^{n} x_i y_i|}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

Donde:

- $|\sum_{i=1}^{n} x_i y_i|$ representa el valor absoluto de la sumatoria del producto punto de los vectores x e y.
- $\sqrt{\sum_{i=1}^{n} x_i^2}$ y $\sqrt{\sum_{i=1}^{n} y_i^2}$ son las normas de los vectores x e y.
- 1 - $\frac{|\sum_{i=1}^{n} x_i y_i|}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$ representa el coseno del ángulo entre los vectores x e y. Se resta 1 para obtener la similitud del coseno.

El algoritmo K-Means permite realizar agrupamientos mediante una combinación de

algoritmos de cálculo estadístico para la ponderación de términos. Esto facilita la agrupación de documentos o textos similares en conjuntos coherentes, estos algoritmos son:

Frecuencia de Términos (tf) para realizar la agrupación por medio del número de apariciones o frecuencia con la que aparece una palabra en una colección de documentos, donde su principio es que entre mayor sea la frecuencia del término en el documento, mayor será su importancia.

$$tf(t,d) = \frac{f_{t,d}}{\sum t' \in d f_{t,d}}$$

Donde t es la palabra que se está evaluando y d un documento específico del corpus; por lo que $f_{t,d}$ es es la frecuencia del término t en el documento d, es decir, cuántas veces aparece el término t en el documento d.

Frecuencia de Término – Frecuencia Inversa de Documento (tf/idf) que determina si un texto es relevante en relación con los términos que más se usan en un documento contra aquellos términos que son muy frecuentes en varios textos que se están considerando en la misma colección de documentos obteniendo así los términos más específicos del dominio que se está procesando.

$$idf(t,D) = log\frac{N}{|d \in D : t \in d|}$$

Donde D es el conjunto de todos los documentos con los que se está trabajando y *N:* número total de documentos en el corpus $N = |D|$

$|d \in D : t \in d|$ número de documentos donde el término t aparece, es decir $tf(t,d) \neq 0$ si el término no está en el corpus, esto dará una división en cero. Por lo tanto, es común ajustar el denominador $1 + |d \in D : t \in d|$ Se define el cálculo de TF/IDF como $tfidf(t,d,D) = tf(t,d) * idf(t,D)$.

El algoritmo K-Means espera como parámetros una colección X y un número de clúster K, para obtener el número de clúster ideal para que se aplican las medidas internas de

inercia.

El algoritmo K-Means requiere como parámetros una colección, denotada como X, y un número de clúster, representado por N. Para determinar el número óptimo de clústeres sobre el cual aplicar las medidas internas de inercia, se utiliza una de las métricas conocidas como la Suma de Cuadrados del Error(SSE). Esta métrica evalúa la dispersión de los datos dentro de los clústeres, ayudando así a identificar la configuración de clústeres que mejor representa la estructura inherente de los datos [37].

$$SSE = \sum_{k=1}^{N} \sum_{\forall x_i \in C_k} ||x_i - \mu_k||^2$$

Donde C_k es el conjunto de instancias del grupo k; μ_k es la media vectorial del cluster k. Y los componentes de μ_k se calculan como:

$$\mu_{kj} = \frac{1}{N_k} \sum_{\forall x_i \in C_k} X_{ij}$$

Donde $N_k = C_k$ es el número de instancias que pertenecen al grupo k.

Adicional a la medida de la Suma de Cuadrados del Error (SSE), el Coeficiente de Silueta, también conocido como Coeficiente de *Silhouette*, se utiliza para ayudar a determinar el número óptimo de clústeres en el contexto de un repositorio de documentos [7] [38]. Este Coeficiente se calcula de la siguiente manera:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

Donde:

- $a(i)$ es el promedio de las disimilitudes (o distancias) de la observación i con las demás observaciones del cluster al que pertenece i.

  Se calcula como:

$$a(i) = \frac{1}{|Cluster(i)| - 1} \sum_{\forall x_j \in Cluster(i)} d(x_i, x_j)$$

Donde $Cluster(i)$ es el conjunto de puntos en el mismo cluster que i y $d(x_i, x_j)$ es la distancia entre los puntos i y j.

- $b(i)$ es la distancia mínima a otro cluster que no es el mismo en el que está la observación i.
  Se calcula como:

$$b(i) = min_{k \neq Cluster(i)} \frac{1}{|Cluster(k)|} \sum_{\forall x_j \in Cluster(k)} d(x_i, x_j)$$

  Donde k representa todos los clusters distintos al cluster al que pertenece i, por lo tanto, $Cluster(k)$ es el conjunto de puntos en cada cluster diferente de i.
  Ese cluster es la segunda mejor opción para i y se lo denomina vecindad de i.

El valor de $s(i)$ puede ser obtenido combinando los valores de a y b como se muestra a continuación:

$$\begin{cases} 1 - \frac{a}{b}, & \text{si } a < b \\ 0, & \text{si } a = b \\ \frac{b}{a} - 1, & \text{si } a > b \end{cases}$$

El coeficiente de *Silhouette* es un valor que varía entre -1 y 1.

1.1.2. Modelado de tópicos

Latent Dirichlet Allocation (LDA) es un modelo bayesiano jerárquico de tres niveles, en el que cada elemento de una colección se modela como una mezcla finita sobre un conjunto subyacente de temas. A su vez, cada tema se modela como una mezcla infinita sobre un conjunto subyacente de probabilidades de temas, donde su principal objetivo es encontrar los temas latentes en un conjunto de documentos y la distribución de probabilidad de las palabras dentro de cada tema. [8].

1.1.3. Sistemas de búsquedas de respuestas

Debido a las grandes cantidades de información que se generan diariamente, han surgido muchos desafíos para procesar y analizar información, sobre todo en información no estructurada, que se puede obtener de diferentes recursos, aquí radica la importancia de las diferentes técnicas de PLN disponibles para obtener datos concisos y precisos.

Dentro de los campos de investigación de PLN se encuentra los SBR, que han tomado mayor relevancia en los últimos años para contestar preguntas realizadas por usuarios en lenguaje natural sobre un conjunto de datos no estructurados, los cuales pueden ser especializados en un tema o pueden ser abiertos; además, pueden estar en diferentes fuentes y formatos, lo cual genera un incremento en el costo computacional. Dentro de este campo de SBR existen diferentes técnicas o modelos de PLN que se han aplicado para poder responder a preguntas, sin embargo, en los últimos años el aprendizaje profundo se ha desarrollado ampliamente, popularizando las redes neuronales que ha permitido que el rendimiento en estos modelos incremente considerablemente. Los SBR ayudan a explotar de mejor manera la extracción de información sobre un conjunto de datos heterogéneos. [4]

Los SBR son una combinación de Recuperación de Información (RI) y PLN enfocados en responder preguntas en lenguaje natural, contenidas en conjuntos de datos o documentos específicos, que a su vez ofrecen un enfoque de respuestas adecuado y preciso. Los SBR tienen muchos desafíos a nivel discurso, sintaxis y semántica en este campo, estas son una de las direcciones de investigación prominentes de las últimas décadas. [3, 35]

Existen dos ramas en los SBR: dominio abierto y dominio restringido o cerrado, este último considera temas o tópicos especializados y busca obtener información precisa sobre un cúmulo de información. Para los SBR de dominio abierto se utilizan conjuntos de datos muy grandes, ya sean de dominio público o no (como Wikipedia, *Common Crawl*, etc.) y que pueden responder preguntas generales como es: *The Stanford Question Answering Dataset* (SQuAD) que utiliza como conjunto de datos artículos de

Wikipedia donde la respuesta a cada pregunta es un segmento de texto, o lapso del pasaje de lectura correspondiente [36]. Hay diferentes tipos de SBR, como son: respuestas Sí/No; preguntas con opción múltiple; soluciones clásicas en Recuperación de Información; soluciones en Big Data; basadas en ontologías; redes neuronales y otros enfoques como gráficos de conocimiento [34]. Cada uno de estos tipos de SBR utilizan diferentes técnicas para dar respuesta a las preguntas realizadas por el usuario.

Las investigaciones de SBR con redes neuronales han tenido un auge mayor en años recientes debido a los resultados obtenidos por estos modelos y por la cantidad de información con la que se cuenta para poder entrenarlos. El uso de Redes neuronales como: *Convolutional Neural Networks* (CNN) [31], *Long Short-Term Memory* (LSTM) [26], *Bidirectional Encoder Representations from Transformers* (BERT) [21], brindan soluciones con arquitecturas utilizadas para SBR en dominio abierto o cerrado.

La incorporación de arquitecturas basadas en BERT en tareas como los Sistemas de Búsqueda de Respuestas (SBR) es fundamental para comprender cómo estas tecnologías han transformado la manera en que las máquinas comprenden y responden a preguntas en lenguaje natural. Estas arquitecturas mejoran la eficiencia y precisión de estos sistemas al comprender mejor el contexto que los enfoques tradicionales.

1.1.4. BERT

BERT es un algoritmo de Google utilizado para PLN que tiene una arquitectura de *Transformers* basada únicamente en mecanismos de atención, prescindiendo por completo de la recurrencia y las convoluciones. Este tipo de arquitecturas se crearon para entrenar representaciones bidireccionales profundas que toman en cuenta el contexto izquierdo como derecho en todas sus capas. BERT puede resolver diferentes tareas dentro del campo PLN como son: clasificación de texto, reconocimiento de entidades nombradas, sistemas de respuestas de preguntas entre otras. [44]

Una de las características que tienen los modelos recientes de PLN es que aprovechan toda la información disponible para preentrenamiento, en específico todos los textos disponibles [9] y esto se ve reflejado en BERT que utiliza dos conjuntos de datos muy

grandes como Wikipedia y Google Book, que en conjunto contienen millones de palabras para preentrenamiento. A partir de BERT se han ido creando variaciones y versiones de este modelo utilizando una arquitectura de *Transformers* para tareas en específico con modelos ya preentrenados, algunos de los modelos que se encuentran son: DistilBERT [39], RoBERTa [33], ELECTRA [28], ALBERT [48], BETO [9], entre otros. Cada una de estas presentaciones viene configurada con diferentes versiones, simplificadas y versiones completas, donde varía el número de codificadores y el número de parámetros. Cada uno de estos modelos tienen una arquitectura que permiten el uso de GPU para mayor velocidad en el entrenamiento. A continuación, se detalla la descripción de dos modelos: DistilBERT y BETO. Estos modelos, que fueron diseñados para reducir la carga de recursos durante el entrenamiento y la evaluación. Además, es importante destacar que BETO está entrenado específicamente para el idioma español, lo que lo convierte en una herramienta especialmente adecuada para tareas que involucran este idioma.

1.1.4.1. DistilBERT

DistilBERT es una de las variantes de BERT, donde su principal característica es que requieren menos recursos para entrenamiento y evaluación de un modelo sin perder rendimiento, DistilBERT es más pequeño, rápido y ligero con un 40 % menos de parámetros que *bert-base-uncased* y 60 % más rápido de entrenar un modelo y conserva el 97 % de rendimiento que proporciona la versión más completa de BERT. [39]

DistilBERT tiene la misma arquitectura general que BERT, pero con algunas modificaciones. La diferencia clave entre estos dos modelos radica en la eliminación de las funciones de *token-type embeddings* y *pooler*, que se utilizan para representar cada token en una secuencia, así como en la capa de red neuronal que genera el contexto de entrada después de que se ha procesado la capa del transformador bidireccional. Además, se redujo a la mitad el número de capas, lo que disminuye la cantidad de parámetros del modelo original de 110 millones a 66 millones. [39]

Para reducir los recursos necesarios en el preentrenamiento de DistilBERT, se aplicó

la técnica de "destilación de conocimiento". Esta metodología implica transferir el conocimiento al reproducir el comportamiento de un modelo más grande y completo (maestro) en un modelo más pequeño y sencillo (alumno), lo que produce resultados cercanos a los del modelo maestro [25].

La metodología establece una transferencia de conocimiento por parte del maestro al alumno, el preentrenamiento del alumno se da con una función de pérdida llamada "*training loss*", que se calcula como:

$$L_{ce} = \sum_i t_i * log(s_i)$$

Donde t_i es la probabilidad estimada por el maestro y s_i por el alumno, de esta manera, se aprovecha la distribución completa del modelo grande (maestro) y posteriormente se pueda medir la transferencia de conocimiento. Adicional y siguiendo la metodología de "destilación de conocimiento"se agrega la función "*softmax-temperature*" que está dada por:

$$pi = \frac{exp(z_i / T)}{\sum_j exp(z_j / T))}$$

En la cual T controla toda la suavidad de salida y z_i es la calificación de la modelo asignada para la clase i. La misma *temperature* se aplica al maestro y alumno en tiempo de entrenamiento, mientras que inferencia, T se fija en 1 para recuperar *softmax* estándar.

El objetivo final del entrenamiento es una combinación lineal de pérdida de destilación L_{ce} con la pérdida de entrenamiento supervisado, en este caso *Masked Language Modeling Loss* L_{mlm} utilizada en BERT, a esto se añade *Cosine embedding loss* L_{cos} que alinea las direcciones de los vectores de estados ocultos del maestro como del alumno.

1.1.4.2. BETO

El idioma español es una de las cinco lenguas más habladas del mundo. Sin embargo, no es fácil encontrar recursos para entrenar y evaluar modelos lingüísticos en este

idioma, en el año 2020 Cañete et al [9] presentan BETO, un modelo BERT pre entrenado sobre un gran corpus en español, este muy similar a *bert-base* en tamaño y fue entrenado con 110M de parámetros, con 12 capas de autoatención, 16 cabezas de atención cada una, además utiliza 1024 como tamaño oculto. BETO integra la técnica "*Dynamic Masking*" en el entrenamiento, que se refiere al uso de diferentes máscaras para la misma frase en el corpus, mientras que el entrenamiento dinámico que usa está a una proporción de 10x, lo que significa que cada oración tiene 10 máscaras diferentes. También considera la técnica "*Whole-Word Masking (WWM)*" de la versión actualizada de BERT, que garantiza que al enmascarar un token en específico, si el token corresponde a una subpalabra en una oración, entonces todos los tokens contiguos que conforman la misma palabra también se enmascaran. [9]

El trabajo presentado por Cañete et al [9] presenta mejores resultados en comparación con otros modelos basados en BERT preentrenados en corpus multilingües para la mayoría de las tareas de la línea de "*benchmark GLUE*".

1.1.5. Hiperparámetros en redes neuronales

Los hiperparámetros se refieren a la configuración empleada en el entrenamiento de redes neuronales, los cuales influyen en el comportamiento y rendimiento de una red neuronal. Estos hiperparámetros son esenciales para ajustar una red neuronal y obtener un mejor rendimiento y resultados óptimos. El proceso de ajuste de estos hiperparámetros es iterativo en función del conjunto de datos que se está evaluando. A continuación, se describen los que se han utilizado para este proyecto.

Época (*epoch*). Al recorrer todos los datos de entrenamiento por la red neuronal para que aprenda de ellos se dice que se realizó una época.

Tamaño por lotes (*batch size*). El número de datos que tiene cada época en paralelo para entrenar el modelo.

Tasa de aprendizaje (*learning rate*). Define el cambio que toma el descenso del gradiente.

Máximo largo de secuencia (*max length*). Longitud máxima de una característica (pregunta y contexto).

Dilución (*dropout*). Ayuda a reducir el sobre ajuste durante el entrenamiento.

Función de activación (*activation function*). Función que permite aprender relaciones no lineales en los datos de entrada.

1.1.6. Métricas de evaluación

Las métricas de evaluación para un Sistema de Respuestas de Preguntas (SBR) son la capacidad de evaluar un modelo de forma cuantitativa en cuanto a precisión y relevancia. Estas métricas son esenciales para determinar qué tan efectivo es un SBR. A continuación, se describen estas métricas, así como las siglas utilizadas en cada una de sus fórmulas:

- *TP* Donde el modelo indicó que eran positivos y acertó.
- *TN* Donde el modelo indicó que eran negativos y acertó.
- *FP* Donde el modelo indicó que eran positivos y se equivocó.
- *FN* Donde el modelo indicó que eran negativos y se equivocó.

F1. Combina las métricas de *Precision* y *Recall* en una sola métrica, lo que hace adecuada esta medida para cuando se requiere un equilibrio entre ambas medidas ya que es la media armónica ente ambas.

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F1 = 2 * \frac{(precision * recall))}{(precision + recall)}$$

Exact Match. Calcula la proporción de respuestas generadas por el modelo que coinciden exactamente con las respuestas de referencia o respuestas esperadas. Para determinar si hay una coincidencia exacta, se comparan las respuestas generadas por el modelo con las respuestas de referencia y se verifica si son idénticas en términos de contenido y formato.

$$Exact\ Match = \frac{Número\ de\ respuestas\ correctas\ exactas}{Número\ total\ de\ preguntas\ del\ conjunto\ de\ datos}$$

1.2. Estado del arte

Las investigaciones con redes neuronales en el ámbito legal o de regulación encontramos a Kim et al. [30] que propone responder a preguntas de sí/no de exámenes legales de abogados japoneses, con soluciones de técnicas de recuperación de información TF-IDF, clasificación SVM y una red neuronal convolucional donde el modelo de vinculación se centra en *Word Embeddings*, similaridad sintáctica e identificación de relaciones de negación.

Por otro lado, Huang et al. [27] plantea un asistente legal en el dominio de las leyes chinas implementando una red de atención interactiva específicamente una LSTM Bidireccional (Bi-LSTM) para calcular representaciones semánticas de la pregunta y la respuesta, lo que permite que los pares de QA estén al tanto de la información de fondo en el dominio legal y aprovecha el conocimiento del dominio de KG legal (*legal Knowledge Graph* (KG)) para enriquecer el aprendizaje.

Por otra parte, Collarana et al. [17] responde a las preguntas mediante un enfoque de dos pasos: selección de párrafos y selección de respuestas, la selección de párrafos corresponde a los más relevantes de los documentos, utilizando un muestreo negativo para minimizar los márgenes entre entidades relacionadas y maximiza los márgenes entre consultas y texto irrelevante, La selección de respuestas, se realiza mediante dos capas principales, la primera (Match-LSTM indica el grado de coincidencia de cada párrafo con el de la pregunta y la segunda (*Answer-Pointer*) determina el intervalo exacto

del párrafo.

Actualmente los modelos PLN que obtienen los mejores resultados en conjunto de datos sin etiquetar, están basados en las arquitecturas de tipo *Transformer* [44], los cuales han revolucionado las diferentes tareas de PLN como: clasificación de texto, sistemas de búsqueda de respuestas, reconocimiento de entidades nombradas, modelado de lenguaje, entre otras; esto se debe a que estos modelos logran aprender la relación sintáctica y semántica de las palabras observando todas las palabras de la oración. Adicional a las redes neuronales de autoatención, estos modelos se han beneficiado de la cantidad de información no estructurada que está disponible, como son miles libros digitales, páginas web como Wikipedia, entre otros recursos, además de la información hay que hacer énfasis a los recursos existentes para el procesamiento en entrenamiento y evaluación de estas redes neuronales.

Investigaciones de SBR con este tipo de arquitectura se encuentran Zhang et al. [46] que combina el modelo BERT y una red neuronal bidireccional GRU uniendo las preguntas y los párrafos en el modelo BERT, para después realizar la extracción de características bidireccionales a través de la capa de red neuronal Bi-GRU. El modelo BERT es utilizado como una capa de *Embeddings* que considera completamente las características de relación de nivel de carácter, nivel de palabra, nivel de oración-carácter y oración a oración para mejorar la representación semántica de la palabra. Por su parte, Bi-GRU calcula la secuencia de entrada en orden y orden inverso para obtener dos capas ocultas con diferentes representaciones, luego pasa el método de costura vectorial y obtiene la representación final de la característica en una capa oculta para finalmente empalmar ambas capas y obtener un vector de representación conjunta para reducir la pérdida de información.

Por otra parte, Kien et al. [29] introduce un codificador de oraciones y uno de párrafos, donde el codificador de oraciones contiene tres capas: *Word Embeddings*, red CNN y red de atención, las cuales ayudan a reflejar la relación semántica, obtener el contexto local alrededor de las palabras para aprender la representación de la oración completa utilizando los token; mientras que el segundo, calcula el peso de la atención de una oración promediando los pesos de atención de las palabras que pertenecen a la misma.

Hasta este momento, no hay disponibles implementaciones de SBR con conjuntos de datos en materia de Prevención de Lavado de Dinero y Financiamiento al Terrorismo (PLD/FT), pero se encuentran investigaciones en PLD con técnicas de Procesamiento de Lenguaje Natural (PLN), entre las cuales se encuentran Chen et al [11] que desarrolla un sistema por lotes distribuido basado en aprendizaje profundo que busca información de dominio público para determinar grupos de clientes con mayor riesgo en posible lavado de dinero mediante noticias negativas en el mundo. El método consiste en tres componentes secuenciales: preprocesamiento de texto, *Word Embeddings* y agrupación por clústeres, para la etapa de preprocesamiento utiliza *Sinica CKIP tokenizer* y *POS tagger* para obtener tokens que se consideran representativos a cada documento, *Word Embeddings Distributed Bag of Words* del modelo *Paragraph Vector* (PV-DBOW) que ignora palabras de contexto para alinear conocimiento empírico en PLD y finamente utiliza *Balanced Iterative Reducing and Clustering Using Hierarchies* (BIRCH) para la creación de clústeres jerárquicos.

Por otra parte, Gavrilkina [6] propone un método de construcción de estructuras conceptuales y terminológicas que se basa en la extracción automática de términos y el análisis de dependencias de su apariencia conjunta utilizando características estadísticas y morfológicas de términos. El análisis de características morfológicas de términos en textos se da mediante el uso de separadores (signos de puntuación), para extraer construcciones con un alto grado de probabilidad de ser frases nominales sustantivas omitiendo adjetivos, sustantivos y preposiciones. Las frases seleccionadas se organizan según el principio de inclusión lexicográfica en forma de diccionario jerárquico dando un enfoque para construir una ontología basada en patrones léxico-sintáctico para la recuperación de información, usando programación genérica y automática para crear modelos.

Finalmente, Zhongfei [47] expone la metodología Descubrimiento de Enlaces basado en Análisis de Correlación (LDCA), utilizando una medida de correlación para determinar la similitud de patrones entre dos elementos del conjunto de datos para inferir la fuerza de su vinculación, además de aplicar lógica difusa para cabida a la imprecisión de la similitud de los patrones. Para realizar este proceso, se analiza y extrae el vector

del historial transacciones de una colección de documentos y se crea una estructura de datos anidados tridimensionales basados en eventos que sigue el principio de vecino más cercano unidireccional. Una vez creado el mapa de transacciones por sujeto se realiza agrupamiento estándar con el algoritmo *K-mean* con espacios euclidianos proyectando el vector del historial de transacciones financieras a variables escalares para formar histogramas y poder reducir el problema de agrupación a segmentación y finalmente obtener la correlación de los individuos con una combinación de una correlación local con una correlación global a partir de sus histogramas.

Dominio	Sistemas de Búsqueda de Respuestas	Modelos	Técnicas utilizadas pre-procesamiento
Legal	A Free Format Legal Question Answering System	Siameses BERT Legal	BM25 y Glove
	A Question Answering System on Regulatory Documents	Match-LSTM y Answer-Pointer	
	AILA: A Question Answering System in the Legal Domain	Bi-LSTM y gráfico de conocimiento legal (LKG)	
	Answering Legal Questions by Learning Neural Attentive Text Representation	Red neuronal CNN y atención	Embeddings
	Applying a Convolutional Neural Network to Legal Question Answering	Red neuronal CNN	TF-IDF y SVM
	Questions and Answers on Legal Texts Based on BERT-BiGRU	BERT, red neuronal bi-direccional Bi-GRU	
	Using Graphs for Shallow Question Answering on Legal Documents	LKG y Dijkstra	
PLD	Applying Data Mining in Investigating Money Laundering Crimes	Descubrimiento de Enlaces basado en Análisis de Correlación (LCDA)	kMeans
	Modeling of Conceptual and Terminological Structures Based on AML/CFT Texts for Solving Problems of Semantic Search	Estructuras conceptuales terminológicas	
	Pluto: A Deep Learning based Watchdog for Anti Money Laundering	Sinica CKIP tokenizer, POS tagger, PV-DBOW, BIRCH	

Cuadro 1.1: Sistemas de Búsquedas de Respuestas con técnicas de PLN.
Fuente: Elaboración propia.

Capítulo 2

Metodología para la construcción de un conjunto de datos en materia de PLD/FT

Capítulo 2.

Metodología en conjuntos de datos

En este capítulo se describe la metodología que se utilizó para la construcción del conjunto de datos, insumo principal de los modelos BERT evaluados en este trabajo y las instituciones encargadas de la regulación en materia de Prevención de Lavado de Dinero y Financiamiento al Terrorismo (PLD/FT). Asimismo, se presentan las técnicas de PLN y los resultados del análisis exploratorio de datos (EDA, por sus siglas en inglés) a los recursos del proyecto.

2.1. Repositorio de archivos

Uno de los objetivos específicos de este proyecto es la creación de un conjunto de datos para un Sistema de Búsqueda de Respuestas (SBR) en materia de PLD/FT, para esto, es necesario primero contar con un repositorio de archivos PDF con los documentos oficiales emitidos por las instituciones que se encargan de regulación en esta materia y que son de dominio público. El repositorio de archivos contiene una lista de documentos que se utiliza para la creación del conjunto de datos, que se basa principalmente en el temario establecido por la Comisión Nacional Bancaria y de Valores (CNBV) para la certificación en materia de PLD/FT, así como en el temario de la certificación del sector de Actividades Vulnerables (AV) que otorga la Unidad de Inteligencia Financiera (UIF) y las 40 Recomendaciones del GAFI. Es importante mencionar que las preguntas no se extrajeron del total de documentos, pero se pone a disposición la documentación oficial referente a esta temática.

Después de la recopilación de documentos, se concentraron en una sola carpeta y se pusieron a disposición en un repositorio público GitHub. La carpeta se utilizó para el procesamiento de información y el análisis exploratorio de datos (EDA, por sus siglas en inglés) del proyecto para obtener estadística descriptiva sobre la información que

sirve como insumo para el entrenamiento y evaluación del modelo.

2.1.1. Recursos de la investigación

Entre las instituciones encargadas de emitir los documentos oficiales en materia de PLD/FT en México se encuentran la CNBV, la UIF, Secretaria de Hacienda y Crédito Público (SHCP) y el Servicio de Administración Tributaria (SAT); en cuanto al ámbito internacional, destacan los documentos emitidos por el Grupo de Acción Financiera Internacional (GAFI), el Grupo de Acción Financiera de Latinoamérica (GAFILAT) y el Consejo de Seguridad de las Naciones Unidas (CSNU), a continuación, se describe la función de cada una de estas.

CNBV. Tiene por objeto supervisar y regular a las entidades integrantes del Sistema Financiero Mexicano y otros Sujetos Supervisados, procurando la estabilidad y correcto funcionamiento de dicho sistema en su conjunto. [12]

En materia de PLD/FT, cuenta con las siguientes facultades:

- Supervisa, a través de visitas de inspección o de acciones de vigilancia, que los Sujetos Supervisados cumplan con lo establecido en las Disposiciones de Carácter General, aplicables en materia de PLD/FT.
- Impone las sanciones administrativas a los referidos Sujetos Supervisados.
- Pone a disposición de los Sujetos Supervisados la Lista de Personas Bloqueadas que se recibe por parte de la SHCP.

UIF. Es la instancia central nacional para:

- Recibir reportes de operaciones financieras y avisos de quienes realizan actividades vulnerables;
- Analizar las operaciones financieras y económicas y otra información relacionada; y
- Diseminar reportes de inteligencia y otros documentos útiles para detectar operaciones probablemente vinculadas con el lavado de dinero (LD) o el financia-

miento al terrorismo (FT), y en su caso, presentar las denuncias correspondientes ante la autoridad competente.

- Emitir la Lista de Personas Bloqueadas

Las principales tareas de la Unidad de Inteligencia Financiera consisten en implementar y dar seguimiento a mecanismos de prevención y detección de actos, omisiones u operaciones, que pudieran favorecer, prestar ayuda, auxilio o cooperación de cualquier especie para la comisión de los siguientes delitos previstos en el Código Penal Federal [43]:

- Operaciones con Recursos de Procedencia Ilícita (Artículo 400 Bis)
- Financiamiento al terrorismo (Artículo 139 Quáter)

SHCP. Es la dependencia del Poder Ejecutivo Federal que tiene como misión proponer, dirigir y controlar la política del Gobierno Federal en materia financiera, fiscal, de gasto, de ingresos y deuda pública.

Parte de sus funciones es emitir normas en materia de PLD/FT, supervisar la debida aplicación de dichas normas, así como de recabar, analizar y diseminarla información a los actores clave en PLD/FT. [42]

SAT. Es un órgano desconcentrado de la SHCP, que tiene la responsabilidad de aplicar la legislación fiscal y aduanera, con el fin de que las personas físicas y morales contribuyan proporcional y equitativamente al gasto público y de fiscalizar a los contribuyentes para que cumplan con las disposiciones tributarias y aduaneras. [12]

Asimismo, está encargado de supervisar, verificar y vigilar el cumplimiento de las obligaciones de aquellas actividades y profesiones que no son de naturaleza financiera pero que representan cierto riesgo para el LD, también llamadas Actividades Vulnerables, y los actos u operaciones para los que existe la restricción de liquidar o pagar, así como de aceptar la liquidación o el pago de acto su operaciones mediante el uso de monedas y billetes en moneda nacional o cualquier otra divisa (Efectivo) y Metales Preciosos. [12]

GAFI. El Grupo de Acción Financiera Internacional (GAFI) es un ente interguberna-

mental conformado por 39 jurisdicciones, que se encarga de fijar estándares y promover la implementación efectiva de medidas legales, regulatorias y operativas para combatir el lavado de activos, el financiamiento del terrorismo y el financiamiento de la proliferación y otras amenazas a la integridad del sistema financiero internacional. Las Recomendaciones del GAFI son reconocidas como el estándar global en materia de prevención y combate al LD/FT/FP. [1]

GAFILAT. Es el Organismo Regional Estilo GAFI de América Latina; se creó en el año 2000 y se compone de 18 países. Entre sus objetivos principales se encuentran el participar en las evaluaciones de sus países miembros de conformidad con los estándares del GAFI, dar seguimiento a los avances reportados por los países miembros, coordinar que los documentos generados por GAFI tengan difusión en la región, así como desarrollar proyectos y cursos en la materia para los países miembros. [23]

CSNU. Tiene la responsabilidad primordial de mantener la paz y la seguridad internacionales. El Consejo de Seguridad tiene 15 miembros y cada miembro tiene un voto. De acuerdo con la Carta, todos los Miembros de la ONU convienen en aceptar y cumplir las decisiones del Consejo de Seguridad. Éste es el único órgano de la ONU cuyas decisiones los Estados Miembros, conforme a la Carta, están obligados a cumplir. [18] Las Resoluciones emitidas por el CSNU obligan a los Estados miembros a aplicar las medidas para restablecer la paz y seguridad internacionales, con fundamento en el Capítulo VII de la CNU titulado “Acción en casos de amenazas a la paz, quebrantamientos de la paz o actos de agresión”. [2]

2.2. Metodología para la creación del conjunto de datos

La Prevención del Lavado de Dinero y Financiamiento al Terrorismo (PLD/FT) es un tópico muy especializado y poco explorado dentro del Procesamiento de Lenguaje Natural (PLN) y aprendizaje profundo (DL, por sus siglas en inglés), así que el reto principal de este proyecto es que no hay un conjunto de datos con el cual se pueda entrenar y evaluar un modelo para un Sistema de Búsqueda de Respuestas (SBR) y poder realizar esta tarea. En este sentido, en este proyecto se crea un conjunto de datos con informa-

ción relevante en PLD/FT.

La metodología empleada en la elaboración de las preguntas se considera desde aspectos básicos como son las siglas y acrónimos utilizados en materia PLD/FT por parte de las instituciones, las autoridades y los organismos, y en la regulación y normatividad; así como preguntas especializadas para los usuarios con mayor experiencia, de esta manera, el modelo se entrenó con información especializada, cabe destacar que la información utilizada se basa en documentos públicos relacionados con la certificación en PLD/FT por parte de la CNBV y/o certificación otorgada por la UIF. A continuación, se muestra la metodología utilizada para crear el conjunto de datos que se utiliza para el entrenamiento y evaluación del modelo BERT, ésta consta de cuatro etapas que se describen de la siguiente manera:

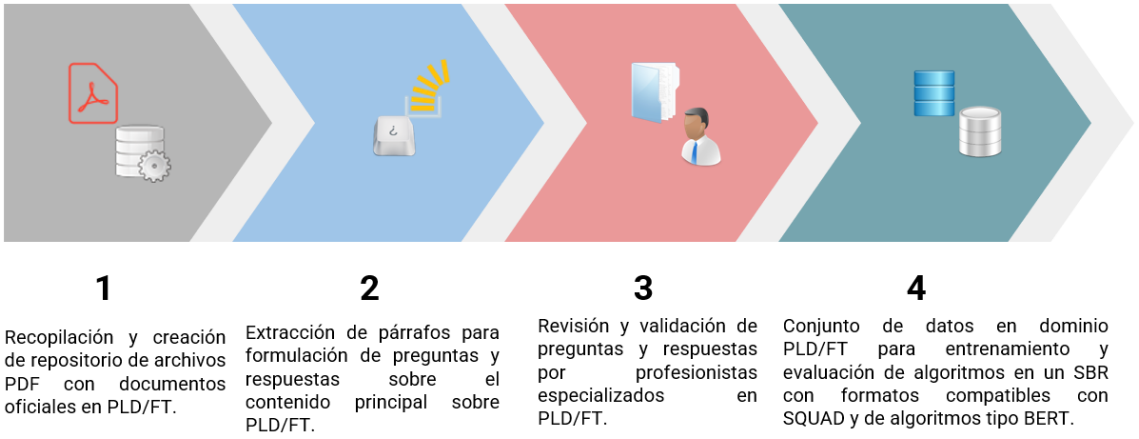


Figura 2.1: Metodología de creación para el conjunto de datos.
Fuente: Elaboración propia.

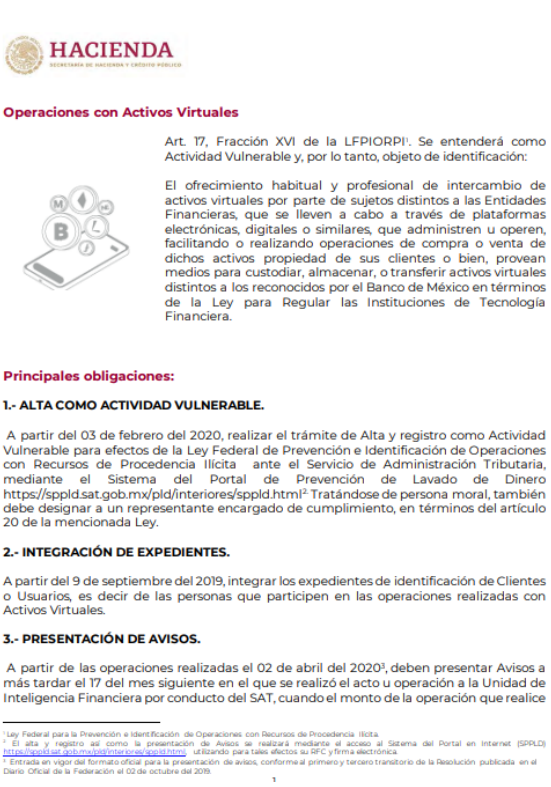
Etapa 1. Recopilación y creación del repositorio de documentos oficiales en PLD/FT.

Como primera etapa se realiza la recopilación de documentos PDF que posteriormente se concentran en el repositorio público del *Especialista Virtual PLD* en GitHub[2], estos archivos tienen la característica de que son emitidos por instituciones oficiales encargadas de esta regulación y que además son de dominio público.

Para el sector no financiero, es decir, las Actividades Vulnerables (AV), se realizó un ajuste a los archivos que ponen a disposición la UIF y el SAT en su portal de internet,

[2]GitHub. Especialista Virtual PLD. Repositorio de archivos con temática PLD/FT. (s.f.). https://github.com/raulrdiaz/Especialista-Virtual-PLD/tree/master/PDF

esta modificación consiste en incorporar un extracto de texto (nombre de la Actividad Vulnerable) en los documentos para poder identificar los umbrales establecidos a cada uno de estos sectores. Lo anterior, debido a que estos documentos siguen una estructura muy similar y genérica en la cual no es posible diferenciar el umbral de cada sector, por lo tanto, se no se pueden agregar preguntas en el entrenamiento y evaluación del modelo. Se realiza este proceso ya que es muy importante conocer cada uno de los umbrales en la que los Sujetos Obligados (SO) tienen la obligación de reportar a la Unidad de Inteligencia Financiera (UIF). Esta incorporación de texto se realiza a los veinte documentos PDF de cada una de las AV.



HACIENDA

Operaciones con Activos Virtuales

Art. 17, Fracción XVI de la LFPIORPI[1]. Se entenderá como Actividad Vulnerable y, por lo tanto, objeto de identificación:

El ofrecimiento habitual y profesional de intercambio de activos virtuales por parte de sujetos distintos a las Entidades Financieras, que se lleven a cabo a través de plataformas electrónicas, digitales o similares, que administren u operen, facilitando o realizando operaciones de compra o venta de dichos activos propiedad de sus clientes o bien, provean medios para custodiar, almacenar, o transferir activos virtuales distintos a los reconocidos por el Banco de México en términos de la Ley para Regular las Instituciones de Tecnología Financiera.

Principales obligaciones:

1.- ALTA COMO ACTIVIDAD VULNERABLE.

A partir del 03 de febrero del 2020, realizar el trámite de Alta y registro como Actividad Vulnerable para efectos de la Ley Federal de Prevención e Identificación de Operaciones con Recursos de Procedencia Ilícita ante el Servicio de Administración Tributaria, mediante el Sistema del Portal de Prevención de Lavado de Dinero https://sppld.sat.gob.mx/pld/interiores/sppld.html[2] Tratándose de persona moral, también debe designar a un representante encargado de cumplimiento, en términos del artículo 20 de la mencionada Ley.

2.- INTEGRACIÓN DE EXPEDIENTES.

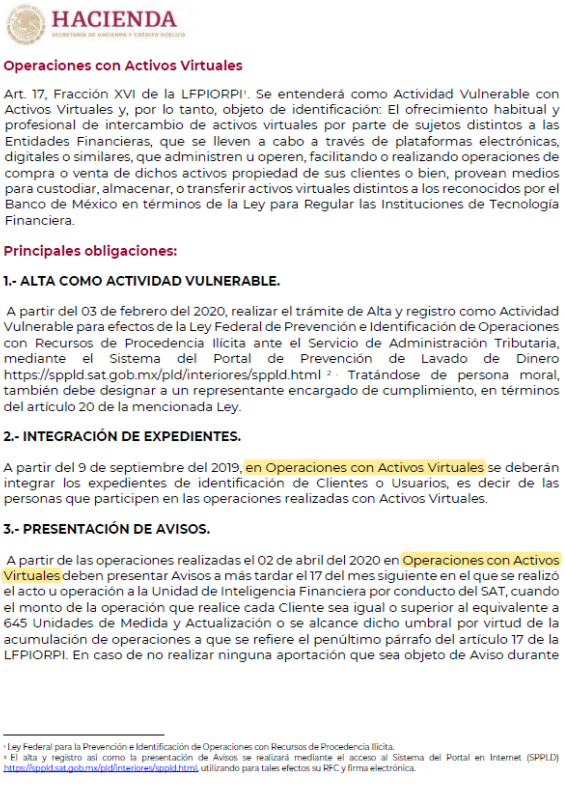
A partir del 9 de septiembre del 2019, integrar los expedientes de identificación de Clientes o Usuarios, es decir de las personas que participen en las operaciones realizadas con Activos Virtuales.

3.- PRESENTACIÓN DE AVISOS.

A partir de las operaciones realizadas el 02 de abril del 2020[3], deben presentar Avisos a más tardar el 17 del mes siguiente en el que se realizó el acto u operación a la Unidad de Inteligencia Financiera por conducto del SAT, cuando el monto de la operación que realice

[1] Ley Federal para la Prevención e Identificación de Operaciones con Recursos de Procedencia Ilícita.
[2] El alta y registro así como la presentación de Avisos se realizará mediante el acceso al Sistema del Portal en Internet (SPPLD) https://sppld.sat.gob.mx/pld/interiores/sppld.html utilizando para tales efectos su RFC y firma electrónica.
[3] Entrada en vigor del formato oficial para la presentación de avisos, conforme al primero y tercero transitorio de la Resolución publicada en el Diario Oficial de la Federación el 02 de octubre del 2019.

Figura 2.2: Formato de Activos Virtuales. Fuente: https://sppld.sat.gob.mx/pld/index.html



HACIENDA

Operaciones con Activos Virtuales

Art. 17, Fracción XVI de la LFPIORPI[1]. Se entenderá como Actividad Vulnerable con Activos Virtuales y, por lo tanto, objeto de identificación: El ofrecimiento habitual y profesional de intercambio de activos virtuales por parte de sujetos distintos a las Entidades Financieras, que se lleven a cabo a través de plataformas electrónicas, digitales o similares, que administren u operen, facilitando o realizando operaciones de compra o venta de dichos activos propiedad de sus clientes o bien, provean medios para custodiar, almacenar, o transferir activos virtuales distintos a los reconocidos por el Banco de México en términos de la Ley para Regular las Instituciones de Tecnología Financiera.

Principales obligaciones:

1.- ALTA COMO ACTIVIDAD VULNERABLE.

A partir del 03 de febrero del 2020, realizar el trámite de Alta y registro como Actividad Vulnerable para efectos de la Ley Federal de Prevención e Identificación de Operaciones con Recursos de Procedencia Ilícita ante el Servicio de Administración Tributaria, mediante el Sistema del Portal de Prevención de Lavado de Dinero https://sppld.sat.gob.mx/pld/interiores/sppld.html [2] Tratándose de persona moral, también debe designar a un representante encargado de cumplimiento, en términos del artículo 20 de la mencionada Ley.

2.- INTEGRACIÓN DE EXPEDIENTES.

A partir del 9 de septiembre del 2019, en Operaciones con Activos Virtuales se deberán integrar los expedientes de identificación de Clientes o Usuarios, es decir de las personas que participen en las operaciones realizadas con Activos Virtuales.

3.- PRESENTACIÓN DE AVISOS.

A partir de las operaciones realizadas el 02 de abril del 2020 en Operaciones con Activos Virtuales deben presentar Avisos a más tardar el 17 del mes siguiente en el que se realizó el acto u operación a la Unidad de Inteligencia Financiera por conducto del SAT, cuando el monto de la operación que realice cada Cliente sea igual o superior al equivalente a 645 Unidades de Medida y Actualización o se alcance dicho umbral por virtud de la acumulación de operaciones a que se refiere el penúltimo párrafo del artículo 17 de la LFPIORPI. En caso de no realizar ninguna aportación que sea objeto de Aviso durante

[1] Ley Federal para la Prevención e Identificación de Operaciones con Recursos de Procedencia Ilícita.
[2] El alta y registro así como la presentación de Avisos se realizará mediante el acceso al Sistema del Portal en Internet (SPPLD) https://sppld.sat.gob.mx/pld/interiores/sppld.html utilizando para tales efectos su RFC y firma electrónica.

Figura 2.3: Formato modificado de Activos Virtuales. Fuente: https://sppld.sat.gob.mx/pld/index.html

En las imágenes anteriores se muestra el documento original y el modificado de la Actividad Vulnerable "Activos Virtuales (AVI)" publicado en el Portal de Prevención de Lavado de Dinero[3]. En la figura 2.3, se remarca el texto "en operaciones con Activos

[3] Servicio de Administración Tributaria. Sistema del Portal en Internet [SPPLD]. (s.f.). https://sppld.sat.gob.mx/pld/interiores/sppld.html

Virtuales" agregado en las secciones "2.-INTEGRACIÓN DE EXPEDIENTES" y "3.- PRESENTACIÓN DE AVISOS" que sirve para identificar el umbral de cada actividad vulnerable.

Para las Disposiciones de Carácter General (DCG), se agregan las diferentes versiones (modificaciones que se han realizado) publicadas en el Diario Oficial de la Federación (DOF) correspondientes a cada de las instituciones del Sector Financiero Mexicano hasta el año 2022, con el objetivo de tener los diferentes cambios realizados a las mismas.

Dentro de este repositorio de documentos también se agregó información relevante que la CNBV y la UIF publican en su portal de internet pero que no está en un documento en formato PDF, por lo que la misma se convierte a este tipo de archivos para concentrarla en este mismo repositorio.

Etapa 2. Extracción de párrafos para formulación de preguntas y respuestas.

La segunda etapa consiste en la extracción de párrafos para la formulación de preguntas y respuestas, éstos se seleccionaron con base a conocimientos básicos y técnicos en materia de Prevención de Lavado de Dinero y Financiamiento al Terrorismo (PLD/FT). Para esta tarea, se toman en cuenta los acrónimos y siglas en la materia, definiciones de conceptos clave, características distintivas entre sujetos obligados, sanciones e infracciones por incumplimiento a las leyes financieras y políticas de aplicación de la normatividad para los perfiles de clientes y usuarios. Cabe mencionar, que esta información se obtuvo de diferentes presentaciones de cursos recopilados para obligaciones mínimas en materia de PLD/FT y de lecturas que la persona que está realizando este proyecto, identifica como relevante. Dentro de los conocimientos básicos y la definición de conceptos clave se plantean preguntas como:

- ¿Qué es el lavado de dinero?
- ¿Qué es la banca múltiple?
- ¿Qué es una Operación Inusual?
- ¿Qué es la Evaluación Nacional de Riesgos?
- ¿Qué es un PEP?

A continuación, se muestran estas preguntas con el formato ya establecido para el conjunto de entrenamiento y evaluación del *Especialista Virtual PLD*.


```
{
    "title": "Glosario de términos portafolio de información CNBV",
    "paragraphs": [{
        "qas": [{
            "id": 79,
            "is_imposible": false,
            "question": "¿Qué es la banca múltiple?",
            "answers": [{
                "text": "Institución financiera de intermediación que recibe fondos en forma de depósito de las personas que poseen
                         excedentes de liquidez, utilizándolos posteriormente para operaciones de préstamo a personas con necesidades
                         de financiación, o para inversiones propias. Presta también servicios de todo tipo relacionados con cualquier
                         actividad realizada en el marco de actuación de un sistema financiero.",
                "answer_start": 17
            }]
        }],
        "context": "Banca múltiple. Institución financiera de intermediación que recibe fondos en forma de depósito de las personas que
                    poseen excedentes de liquidez, utilizándolos posteriormente para operaciones de préstamo a personas con necesidades
                    de financiación, o para inversiones propias. Presta también servicios de todo tipo relacionados con cualquier actividad
                    realizada en el marco de actuación de un sistema financiero."
    }]
}
```


Figura 2.4: Pregunta 79 del conjunto de datos.
Fuente: Elaboración propia.


```
{
    "title": "Lavado de Dinero",
    "paragraphs": [{
        "qas": [{
            "id": 398,
            "is_imposible": false,
            "question": "¿Qué es el lavado de dinero?",
            "answers": [{
                "text": "es el proceso a través del cual es encubierto el origen de los fondos generados mediante el ejercicio
                         de algunas actividades ilegales (siendo las más comunes, tráfico de drogas o estupefacientes, contrabando
                         de armas, corrupción, fraude, trata de personas, prostitución, extorsión, piratería, evasión fiscal y
                         terrorismo).",
                "answer_start": 21
            }]
        }],
        "context": "El lavado de dinero es el proceso a través del cual es encubierto el origen de los fondos generados mediante
                    el ejercicio de algunas actividades ilegales (siendo las más comunes, tráfico de drogas o estupefacientes,
                    contrabando de armas, corrupción, fraude, trata de personas, prostitución, extorsión, piratería, evasión fiscal y terrorismo).
                    El objetivo de la operación, que generalmente se realiza en varios niveles, consiste en hacer que los fondos o activos
                    obtenidos a través de actividades ilícitas aparezcan como el fruto de actividades legítimas y circulen sin problema
                    en el sistema financiero"
    }]
}
```


Figura 2.5: Pregunta 398 del conjunto de datos.
Fuente: Elaboración propia.


```
{
    "title": "Guía para la elaboración de una metodología de evaluación de riesgos en materia de prevención de operaciones con
              recursos de procedencia ilícita y financiamiento al terrorismo",
    "paragraphs": [{
        "qas": [{
            "id": 125,
            "is_imposible": false,
            "question": "¿Qué es la evaluación nacional de riesgos?",
            "answers": [{
                "text": "ejercicio de autoevaluación que realiza el gobierno mexicano con la finalidad de orientar la política de prevención y
                         combate de los riesgos que implican la comisión de los delitos de lavado de dinero, financiamiento al terrorismo, así
                         como a las conductas de financiamiento a la proliferación de armas de destrucción masiva, de manera que se asignen los
                         recursos a la mitigación de los mayores riesgos identificados",
                "answer_start": 33
            }]
        }],
        "context": "Evaluación Nacional de Riesgos, ejercicio de autoevaluación que realiza el gobierno mexicano con la finalidad de orientar la
                    política de prevención y combate de los riesgos que implican la comisión de los delitos de lavado de dinero, financiamiento al
                    terrorismo, así como a las conductas de financiamiento a la proliferación de armas de destrucción masiva, de manera que se asignen
                    los recursos a la mitigación de los mayores riesgos identificados"
    }]
}
```


Figura 2.6: Pregunta 125 del conjunto de datos.
Fuente: Elaboración propia.

```
{
    "title": "Disposiciones de carácter general instituciones de crédito 2021",
    "paragraphs": [{
        "qas": [{
            "id": 99,
            "is_imposible": false,
            "question": "¿Qué es una operación inusual?",
            "answers": [{
                "text": "Operación, actividad, conducta o comportamiento de un Cliente que no concuerde con los antecedentes o actividad
                         conocida por la Entidad o declarada a esta, o con el perfil transaccional inicial o habitual de dicho Cliente,
                         en función al origen o destino de los recursos, así como al monto, frecuencia, tipo o naturaleza de la Operación
                         de que se trate, sin que exista una justificación razonable para dicha Operación",
                "answer_start": 25
            }]
        }],
        "context": "Operación Inusual, a la Operación, actividad, conducta o comportamiento de un Cliente que no concuerde con los antecedentes
                    o actividad conocida por la Entidad o declarada a esta, o con el perfil transaccional inicial o habitual de dicho Cliente,
                    en función al origen o destino de los recursos, así como al monto, frecuencia, tipo o naturaleza de la Operación de que se
                    trate, sin que exista una justificación razonable para dicha Operación, actividad, conducta o comportamiento, o bien, aquella
                    Operación, actividad, conducta o comportamiento que un Cliente realice o pretenda realizar con la Entidad en la que, por
                    cualquier causa, esta considere que los recursos correspondientes pudieran ubicarse en alguno de los supuestos previstos en el
                    artículo 139 Quáter o 400 Bis del Código Penal Federal;"
    }]
}
```


Figura 2.7: Pregunta 99 del conjunto de datos.
Fuente: Elaboración propia.


```
{
    "title": "Guía para la prevención y detección de operaciones con recursos de procedencia ilícita en el sistema
              financiero derivadas de actos de corrupción",
    "paragraphs": [{
        "qas": [{
            "id": 146,
            "is_imposible": false,
            "question": "¿Qué es un PEP?",
            "answers": [{
                "text": "a la Persona Políticamente Expuesta definida como aquel individuo que desempeña o ha desempeñado funciones públicas destacadas
                         en un país extranjero o en territorio nacional, considerando entre otros, a los jefes de estado o de gobierno, líderes políticos,
                         funcionarios gubernamentales, judiciales o militares de alta jerarquía, altos ejecutivos de empresas estatales o funcionarios o miembros
                         importantes de partidos políticos y organizaciones internacionales; entendidas como aquellas entidades establecidas mediante acuerdos
                         políticos oficiales entre estados, los cuales tienen el estatus de tratados internacionales; cuya existencia es reconocida por la Ley
                         en sus respectivos estados miembros, y no son tratadas como unidades institucionales residentes de los países en los que están ubicadas.
                         Adicionalmente, la Secretaría en ejercicio de sus atribuciones, podrá hacer del conocimiento de los Sujetos Supervisados, Datos que
                         permitan identificar en lo individual de dichas personas.",
                "answer_start": 6
            }]
        }],
        "context": "PEP, a la Persona Políticamente Expuesta definida como aquel individuo que desempeña o ha desempeñado funciones públicas destacadas en un
                    país extranjero o en territorio nacional, considerando entre otros, a los jefes de estado o de gobierno, líderes políticos, funcionarios
                    gubernamentales, judiciales o militares de alta jerarquía, altos ejecutivos de empresas estatales o funcionarios o miembros importantes de
                    partidos políticos y organizaciones internacionales; entendidas como aquellas entidades establecidas mediante acuerdos políticos oficiales
                    entre estados, los cuales tienen el estatus de tratados internacionales; cuya existencia es reconocida por la Ley en sus respectivos estados
                    miembros, y no son tratadas como unidades institucionales residentes de los países en los que están ubicadas. Adicionalmente, la Secretaría
                    en ejercicio de sus atribuciones, podrá hacer del conocimiento de los Sujetos Supervisados, Datos que permitan identificar en lo
                    individual de dichas personas."
    }]
}
```


Figura 2.8: Pregunta 146 del conjunto de datos.
Fuente: Elaboración propia.

Adicionalmente a los conocimientos básicos, los técnicos ayudan a comprender cómo se aplica la normatividad en materia de PLD/FT, como son los diferentes umbrales y periodos para la presentación de reportes como: Formato de Transferencias Internacionales (FTI), Formato de Dólares en Efectivo (FDE), Cheques de Caja (CHC), Reportes de Operaciones Inusuales (ROI), Reporte de Operaciones Relevantes (ROR) y Reporte de Operaciones Internas Preocupantes (ROIP), entro otros. Asimismo, se incluyen, artículos de las diferentes leyes como: Instituciones de Crédito, Casas de Cambio, Almacenes Generales de Depósito, Instituciones de Tecnología Financiera, entre otras, con las cuales se rigen las Instituciones Financieras (IF). Es relevante señalar que el conocimiento técnico en PLD/FT se enfoca en todas las reglas de negocio de la parte operacional en las instituciones para su regulación, por lo que es muy importante que tanto analistas, auditores, oficiales de cumplimiento y demás profesionales en materia

PLD/FT tengan conocimiento. Las preguntas sobre conocimientos técnicos son muy específicas como se muestran a continuación:

- ¿Cuál es el monto para envío de reportes de Transferencias Internacionales de Fondos para clientes y usuarios?
- ¿Cuál es el plazo para envío de Reportes de Operaciones Inusuales?
- ¿Cuál es el periodo de tiempo para evaluar el cambio de perfil transaccional de un cliente en ITF?
- ¿Cuál es el monto diario de compra y recepción del pago de servicios o transferencias o situación de fondos en Casas de Cambio para personas físicas?
- ¿Cuándo se presentan los Reportes de Operaciones Relevantes los Centros Cambiarios?
- ¿Cuál es el monto para establecer mecanismos para dar seguimiento en operaciones en efectivo con moneda extranjera por parte de los usuarios como personas físicas en Transmisores de dinero?

A continuación, se muestran las mismas preguntas con el formato ya establecido para el conjunto de entrenamiento y evaluación del *Especialista Virtual PLD*.


```
{
    "title": "DCG Instituciones de crédito",
    "paragraphs": [{
        "qas": [{
            "id": 571,
            "is_imposible": false,
            "question": "¿Cuál es el monto para envío de reportes de transferencias internacionales de fondos para clientes y usuarios?",
            "answers": [{
                "text": "por un monto igual o superior a mil dólares de los Estados Unidos de América o su equivalente en la moneda extranjera en que se realice."
                "answer_start": 358
            }]
        }],
        "context": "Las Entidades deberán remitir mensualmente a la Secretaría, por conducto de la Comisión, a más tardar dentro de los quince días hábiles
                    siguientes al último día hábil del mes inmediato anterior, un reporte por cada transferencia internacional de fondos que, en lo individual,
                    haya recibido o enviado cualquiera de sus Clientes o Usuarios durante dicho mes, por un monto igual o superior a mil dólares de los Estados
                    Unidos de América o su equivalente en la moneda extranjera en que se realice."
    }]
}
```


Figura 2.9: Pregunta 571 del conjunto de datos.
Fuente: Elaboración propia.

```
{
    "title": "DCG Instituciones de crédito",
    "paragraphs": [{
        "qas": [{
            "id": 573,
            "is_imposible": false,
            "question": "¿Cuál es el plazo para envío de reportes de operaciones inusuales?",
            "answers": [{
                "text": "dentro de los tres días hábiles siguientes contados a partir de que concluya la sesión del Comité que la dictamine como tal",
                "answer_start": 147
            }]
        }],
        "context": "Por cada Operación Inusual que detecte una Entidad, esta deberá remitir a la Secretaría, por conducto de la Comisión, el reporte
                    correspondiente, dentro de los tres días hábiles siguientes contados a partir de que concluya la sesión del Comité que la dictamine
                    como tal. Para efectos de llevar a cabo el dictamen en cuestión, la Entidad a través de su Comité, contará con un periodo que no
                    excederá de sesenta días naturales contados a partir de que se genere la alerta por medio de su sistema, modelo, proceso o por el
                    empleado de la Entidad, lo que ocurra primero."
    }]
}
```


Figura 2.10: Pregunta 573 del conjunto de datos.
Fuente: Elaboración propia.


```
{
    "title": "DCG ITF",
    "paragraphs": [{
        "qas": [{
            "id": 567,
            "is_imposible": false,
            "question": "¿Cuál es el periodo para evaluar el cambio de perfil transaccional de un cliente en ITF?",
            "answers": [{
                "text": "al menos cada seis meses",
                "answer_start": 32
            }]
        }],
        "context": "Las ITF deberán llevar a cabo, al menos cada seis meses, la evaluación del perfil transaccional de su Cliente, a fin de determinar
                    si resulta o no necesario modificarlo. Las evaluaciones se realizarán sobre aquellos Clientes cuya celebración de contrato se hubiere
                    realizado al menos con seis meses de anticipación a la evaluación correspondiente."
    }]
}
```


Figura 2.11: Pregunta 567 del conjunto de datos.
Fuente: Elaboración propia.


```
{
    "title": "DCG_Compiladas_Casas_de_Cambio 2022",
    "paragraphs": [{
        "qas": [{
            "id": 513,
            "is_imposible": false,
            "question": "¿Cuál es el monto diario de compra y recepción del pago de servicios o transferencias o situación de fondos en
                         Casas de Cambio para personas físicas?",
            "answers": [{
                "text": "mayor a trescientos dólares de los Estados Unidos de América.",
                "answer_start": 310
            }]
        }],
        "context": "Las Casas de Cambio deberán abstenerse de recibir de Usuarios personas físicas dólares de los Estados Unidos de América
                    en efectivo para la realización de operaciones individuales diarias de compra y recepción del pago de servicios o
                    transferencias o situación de fondos, por un monto en conjunto por Usuario mayor a trescientos dólares de los Estados
                    Unidos de América."
    }]
}
```


Figura 2.12: Pregunta 513 del conjunto de datos.
Fuente: Elaboración propia.


```
{
    "title": "DCG de Centros cambiarios",
    "paragraphs": [{
        "qas": [{
            "id": 522,
            "is_imposible": false,
            "question": "¿Cuándo se presentan los reportes de operaciones relevantes los Centros Cambiarios?",
            "answers": [{
                "text": "dentro de los últimos diez días hábiles de los meses de enero, abril, julio y octubre de cada año",
                "answer_start": 86
            }]
        }],
        "context": "Los Centros Cambiarios deberán remitir a la Secretaría, por conducto de la Comisión, dentro de los últimos diez días
                    hábiles de los meses de enero, abril, julio y octubre de cada año, a través de medios electrónicos y en el formato
                    oficial que para tal efecto expida la Secretaría, conforme a los términos y especificaciones señalados por esta última,
                    un reporte por todas las Operaciones Relevantes"
    }]
}
```


Figura 2.13: Pregunta 522 del conjunto de datos.
Fuente: Elaboración propia.

```
{
    "title": "DCG Transmisor de dinero",
    "paragraphs": [{
        "qas": [{
            "id": 530,
            "is_imposible": false,
            "question": "¿Cuál es el monto para establecer mecanismos para dar seguimiento en operaciones en efectivo con moneda
                        extranjera por parte de los usuarios como personas físicas en Transmisores de dinero?",
            "answers": [{
                "text": "por montos iguales o superiores a quinientos dólares de los Estados Unidos de América o su equivalente en
                        la moneda nacional o moneda extranjera de que se trate.",
                "answer_start": 197
            }]
        }],
        "context": "Los Transmisores de Dinero deberán establecer mecanismos para dar seguimiento y, en su caso, agrupar las Operaciones
                    en moneda extranjera que, en lo individual, realicen sus Usuarios en efectivo, por montos iguales o superiores a
                    quinientos dólares de los Estados Unidos de América o su equivalente en la moneda nacional o moneda extranjera
                    de que se trate."
    }]
}
```


Figura 2.14: Pregunta 530 del conjunto de datos.
Fuente: Elaboración propia.

Con el fin de complementar los conocimientos a nivel nacional, se toma en cuenta la parte internacional donde México pertenece a la red global GAFI/FATF, están los estándares definidos internacionalmente por el GAFI sobre la lucha contra el lavado de activos y el financiamiento al terrorismo, de tal manera que se pueda promover la implementación de medidas legales efectivas, regulatorias y operativas por parte de cada una de la países que conforman el grupo, estos estándares los podemos encontrar en las 40 Recomendaciones donde se definen un esquema muy completo para que los diferentes países puedan realizar una correcta prevención en LD/FT. Las preguntas en la esfera internacional están enfocadas en la 40 Recomendaciones del GAFI y sus Notas Interpretativas:

- ¿Para qué se aplica un Enfoque Basado en Riesgo (EBR)?
- ¿Qué recomendación hace énfasis a obligar a los países a incluir los delitos de financiamiento del terrorismo como delitos determinantes para el lavado de activos?
- ¿Cuál recomendación requiere que se incluyan los Reportes de Transacciones Sospechosas?
- ¿Cómo se tipifica el Lavado de Activos?
- ¿Qué es el financiamiento del terrorismo?

A continuación, se muestran las mismas preguntas con el formato ya establecido para el conjunto de entrenamiento y evaluación del *Especialista Virtual PLD*.

```
{
    "title": "RECOMENDACIÓN 1. Evaluaciónde riesgos y aplicación de unenfoque basado en riesgo",
    "paragraphs": [{
        "qas": [{
            "id": 487,
            "is_imposible": false,
            "question": "¿Para qué se aplica un enfoque basado en riesgo EBR?",
            "answers": [{
                "text": "a fin de asegurar que lasmedidas para prevenir o mitigar el lavado de activos y el financiamiento
                         del terrorismo sean proporcionales a losriesgos identificados.",
                "answer_start": 408
            }]
        }],
        "context": "Los países deben identificar, evaluar y entender sus riesgos de lavado de activos/financiamiento del terrorismo,
                    y deben tomar acción, incluyendo la designación de una autoridad o mecanismo para coordinar acciones paraevaluar
                    los riesgos, y aplicar recursos encaminados a asegurar que se mitiguen eficazmente los riesgos. Conbase en esa
                    evaluación, los países deben aplicar un enfoque basado en riesgo (EBR) a fin de asegurar que lasmedidas para
                    prevenir o mitigar el lavado de activos y el financiamiento del terrorismo sean proporcionales a losriesgos
                    identificados. Este enfoque debe constituir un fundamento esencial para la asignación eficaz de recursosen todo el
                    régimen antilavado de activos y contra el financiamiento del terrorismo (ALA/CFT) y laimplementación de medidas
                    basadas en riesgo en todas las Recomendaciones del GAFI. Cuando los países identifiquen riesgos mayores,
                    éstos deben asegurar que sus respectivos regímenes ALA/CFT abordenadecuadamente tales riesgos. Cuando los países
                    identifiquen riesgos menores, éstos pueden optar por permitirmedidas simplificadas para algunas Recomendaciones
                    del GAFI bajo determinadas condiciones."
    }]
}
```


Figura 2.15: Pregunta 487 del conjunto de datos.
Fuente: Elaboración propia.


```
{
    "title": "RECOMENDACIÓN 5. Delito definanciamiento del terrorismo",
    "paragraphs": [{
        "qas": [{
            "id": 494,
            "is_imposible": false,
            "question": "¿Qué recomendación hace énfasis a obligar a los países a incluir los delitos de financiamiento
                         del terrorismo como delitos determinantes para el lavado de activos?",
            "answers": [{
                "text": "Recomendación 5",
                "answer_start": 4
            }]
        }],
        "context": "La Recomendación 5 fue desarrollada con el objetivo de asegurar que los países contaran con la capacidad
                    legal para procesar y aplicar sanciones penales a las personas que financien el terrorismo.Dada la estrecha
                    conexión entre el terrorismo internacional y, entre otros, el lavado de activos, otro objetivo de la
                    Recomendación 5 es hacer énfasis en este vínculo al obligar a los países a incluir los delitos de
                    financiamiento del terrorismo como delitos determinantes para el lavado de activos."
    }]
}
```


Figura 2.16: Pregunta 494 del conjunto de datos.
Fuente: Elaboración propia.


```
{
    "title": "Estándares internacionales sobre la lucha contra el lavado de activos, el financiamiento del terrorismo,
              y el financiamiento de la proliferación de armas de destrucción masiva",
    "paragraphs": [{
        "qas": [{
            "id": 287,
            "is_imposible": false,
            "question": "¿Cuál recomendación requiere que se incluyan los reportes de transacciones sospechosas?",
            "answers": [{
                "text": "Recomendación 20 y 23",
                "answer_start": 210
            }]
        }],
        "context": "La UIF sirve como agencia central para la recepción de la información revelada por los sujetos obligados.
                    Como mínimo, esta información debe incluir los reportes de transacciones sospechosas, como requiere la
                    Recomendación 20 y 23, y debe incluir otra información que requerida por la legislación nacional (como
                    los reportes de transacciones en efectivo, los reportes de transferencias electrónicas y otras
                    declaraciones/revelaciones basadas en el umbral)."
    }]
}
```


Figura 2.17: Pregunta 287 del conjunto de datos.
Fuente: Elaboración propia.

```
{
    "title": "Estándares internacionales sobre la lucha contra el lavado de activos, el financiamiento del terrorismo,
              y el financiamiento de la proliferación de armas de destrucción masiva",
    "paragraphs": [{
        "qas": [{
            "id": 211,
            "is_imposible": false,
            "question": "¿Cómo se tipifica el lavado de activos?",
            "answers": [{
                "text": "como delito de acuerdo con la Convención de Viena y la Convención de Palermo (véase el Artículo 3(1) (b) y (c)
                         de la Convención de Viena y el Artículo 6(1) de la Convención de Palermo).",
                "answer_start": 39
            }]
        }],
        "context": "El lavado de activos debe tipificarse como delito de acuerdo con la Convención de Viena y la Convención de Palermo
                    (véase el Artículo 3(1) (b) y (c) de la Convención de Viena y el Artículo 6(1) de la Convención de Palermo)."
    }]
}
```


Figura 2.18: Pregunta 211 del conjunto de datos.
Fuente: Elaboración propia.


```
{
    "title": "Estándares internacionales sobre la lucha contra el lavado de activos, el financiamiento del terrorismo,
              y el financiamiento de la proliferación de armas de destrucción masiva",
    "paragraphs": [{
        "qas": [{
            "id": 397,
            "is_imposible": false,
            "question": "¿Qué es el financiamiento del terrorismo?",
            "answers": [{
                "text": "es el financiamiento de actos terroristas y de terroristas y organizaciones terroristas",
                "answer_start": 31
            }]
        }],
        "context": "Financiamiento del terrorismo es el financiamiento de actos terroristas y de terroristas y organizaciones terroristas"
    }]
}
```


Figura 2.19: Pregunta 397 del conjunto de datos.
Fuente: Elaboración propia.

Etapa 3. Revisión y validación del conjunto de datos por profesionistas especializados en PLD/FT.

La tercera etapa de esta metodología consiste en la revisión y validación del conjunto de datos por profesionales especializados en Prevención de Lavado de Dinero y Financiamiento al Terrorismo (PLD/FT). Este proceso de revisión se llevó a cabo por una persona experta en esta temática. Si bien hubiera sido esencial contar con más profesionales especializados en PLD/FT para lograr una revisión precisa y garantizar la calidad de los datos, lamentablemente, debido a la falta de disponibilidad de más personas con experiencia en esta temática, la revisión conjunta de datos se vio limitado en este aspecto. No obstante, se trabajó en varias etapas para dar lugar a la retroalimentación y realizar las modificaciones necesarias. Las preguntas se crearon en base a presentaciones sobre las obligaciones mínimas en materia de PLD/FT y a lecturas relevantes en la materia.

Las preguntas establecidas en este conjunto de datos se validaron por una profesio-

nista especializada dentro de esta regulación, por lo que una vez creada la primera versión se enviaron a revisión y se solicitó retroalimentación con el fin de poder incorporar preguntas sobre aspectos que se consideran indispensables o aquellas preguntas que deberían ser descartadas por redundancia o por la poca relevancia que representan en el tema. Este proceso iterativo continuó hasta la versión final del conjunto de datos con dominio PLD/FT.

Etapa 4. Conjunto de datos en dominio PLD/FT con formatos compatibles con librerías SQuAD y algoritmos BERT.

Finalmente, la cuarta etapa de la metodología es la definición final del conjunto de datos con dominio PLD/FT, insumo principal con el que se entrena y evalúa el rendimiento con modelos BERT para el sistema de preguntas de respuestas del *Especialista Virtual PLD*. El conjunto de datos está establecido con formatos compatibles con modelos SQuAD y tipo BERT para ponerlo a disposición en el repositorio GitHub y se puedan realizar implementaciones propias en un dominio poco conocido. El formato queda definido con los campos:

- *id*: identificador del número de pregunta.
- *title*: título del documento de donde se formuló la pregunta y se extrae la respuesta.
- *context*: párrafo o texto a partir del cual se formula la pregunta.
- *question*: Pregunta a responder.
- *answer_start*: índice inicial de la respuesta en el texto o párrafo.
- *is_imposible*: indica si la pregunta se puede responder correctamente desde el contexto.

A continuación, se muestra la estructura con los campos mencionados anteriormente con el formato JSON para entrenamiento de modelos SQuAD y BERT.

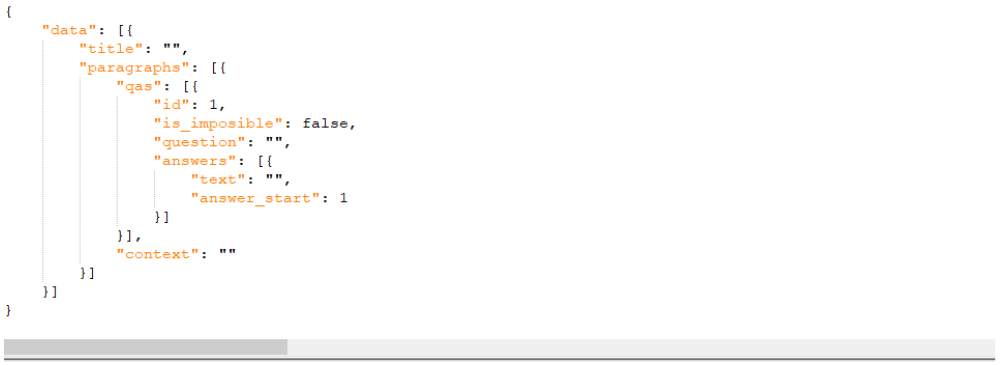


Figura 2.20: Formato JSON para modelos SQuAD y BERT.
Fuente: Elaboración propia.

2.2.1. Análisis exploratorio de datos (EDA)

El análisis exploratorio de datos (EDA, por sus siglas en inglés) EDA se realizó con los recursos disponibles de este proyecto, el repositorio de documentos PDF concentrado en la carpeta pública de GitHub y sobre las preguntas y respuestas definidas en el conjunto de datos para obtener datos estadísticos y comprender mejor la información que se va a procesar, entrenar y evaluar.

Los algoritmos que se utilizaron para el EDA en el repositorio de archivos son: K-Means combinados con algoritmos de cálculo de relevancia en términos como TF y TF/IDF y modelado de tópicos *Latent Dirichlet Allocation* (LDA) para observar cuántas tópicos se pueden identificar en la temática de PLD/FT y el EDA que se aplica sobre las preguntas y respuestas del conjunto de datos son para obtener estadística descriptiva sobre longitudes de párrafos, preguntas, y respuestas así como la distribución de preguntas.

2.2.2. Análisis exploratorio de datos aplicado al repositorio de archivos

La normalización de textos es un paso esencial para cualquier tarea PLN, con este proceso se generan mejores análisis de datos debido a que se realiza una transformación del corpus y “limpieza de datos” para tener una versión más manejable al procesar la información. Al realizar el EDA sobre el repositorio de archivos se obtiene la información de cada uno de los documentos PDF para su procesamiento, esta colección de

documentos tiene la particularidad de ser documentos oficiales, por lo tanto, la calidad de los textos en los documentos es muy buena y no es necesario aplicar una normalización como a textos extraídos de redes sociales. La normalización aplicada al repositorio de archivos se tomó como referencia con base a algunas las funciones de preprocesamiento utilizados en microTC, así como los algoritmos de medidas de ponderación de términos [40]. Estas funciones de preprocesamiento se describen a continuación:

1. **Eliminación de saltos de línea**. Los documentos PDF cuentan con una estructura formal, por tanto, tienen numerosos saltos de línea para dividir secciones o capítulos.
2. **Eliminación de signos de puntuación**. Los signos de puntuación no aportan en la realización de análisis de texto, por lo que fueron eliminados.
3. **Eliminación de acentos**. Todas las palabras con acentos y cualquier signo ortográfico se sustituyeron por el carácter sin el mismo.
4. **Eliminación de espacios en blanco**. Se remplazan dos o más espacios en blanco seguidos por un solo espacio.
5. **Eliminación de palabras vacías (stopwords)**. Se eliminan las palabras vacías que no aportan nada al contexto de la temática, pero son necesarias e indispensables al escribir un texto.
6. **Minúsculas**. Se realiza una transformación de texto en todo el corpus para que cada carácter esté en minúscula.
7. **Lematización**. Reduce las variantes morfológicas de las formas de una palabra a raíces comunes o lexemas.

El repositorio de archivos del *Especialista Virtual PLD* se compone de 262 documentos PDF oficiales y de dominio público emitidos por las diferentes instituciones encargadas de la regulación en materia en PLD/FT entre la que se encuentra, las disposiciones de carácter general, evaluaciones nacionales de riesgo, documentos relacionados al combate contra al financiamiento del terrorismo, lineamientos, tipologías, leyes, reglamentos, así como los estándares nacionales e internacionales que existen. Por su lado, el conjunto de datos para entrenamiento y evaluación de rendimiento se conforma de 703 preguntas cuya respuesta es un segmento de un párrafo dentro de los documentos del repositorio de archivos. Al realizar el EDA se obtiene los siguientes

datos estadísticos:

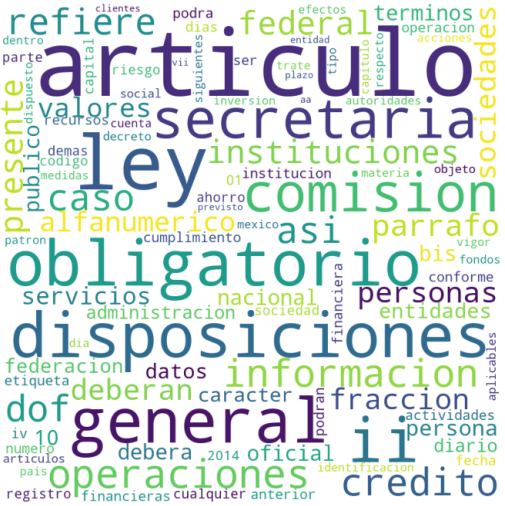


Figura 2.21: Nube de palabras de los cien términos más comunes.
Fuente: Elaboración propia.

Vocabulario único	Vocabulario único sin palabras vacías
34,157	33,880

Cuadro 2.2: Vocabulario único en temática en PLD/FT.
Fuente: Elaboración propia.

2.2.3. Clusterización de documentos

Para la clusterización de documentos se realiza con el algoritmo de aprendizaje no supervisado K-Means que agrupa en k grupos a partir de sus características donde cada punto es un párrafo dentro de un documento, esta clusterización se realiza de dos maneras, la primera solo aplicando una parte de la normalización mencionado anteriormente como son: la eliminación de saltos de línea, signos de puntuación, y espacios en blanco, así como conversión del texto en minúsculas, para la segunda manera, adicional a lo mencionado se aplica la eliminación de *stopwords* y se utiliza el proceso de lematización, este último puede ser más o menos útil dependiendo del idioma con el que se esté trabajando, y como el presente trabajo los textos están en el idioma español, este idioma cuenta con una gran cantidad de morfemas por palabra, por lo que es útil en la clusterización de documentos.

A continuación, se muestran los resultados de la clusterización de párrafos con K-

Means que se realiza sobre la toda la información contenida en el repositorio de documentos con dominio PLD/FT.

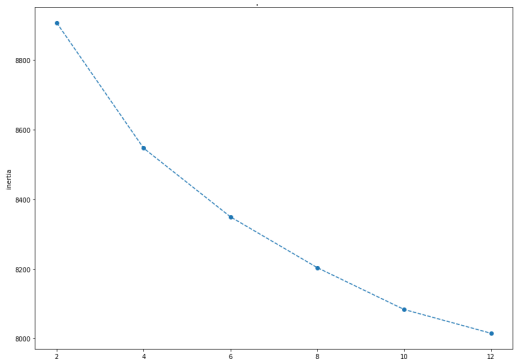


Figura 2.22: Medida Inercia con normalización aplicada.
Fuente: Elaboración propia.

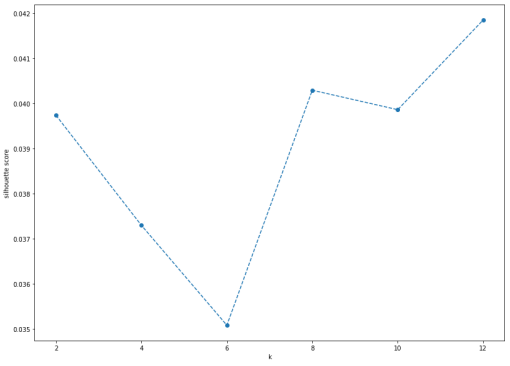


Figura 2.23: Medida Coeficiente de *Silhouette* con normalización aplicada.
Fuente: Elaboración propia.

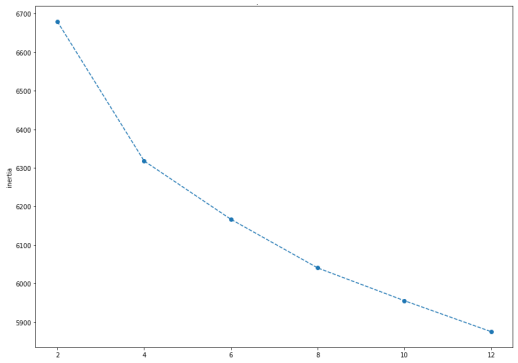


Figura 2.24: Medida Inercia sin normalización aplicada.
Fuente: Elaboración propia.

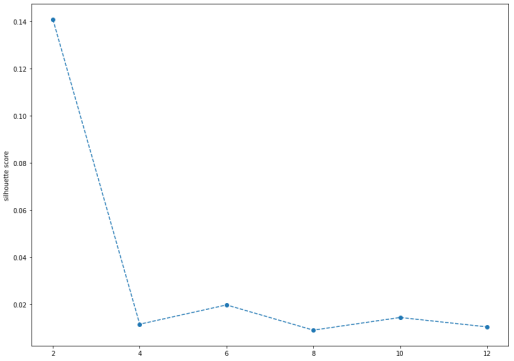


Figura 2.25: Medida Coeficiente de *Silhouette* sin normalización aplicada.
Fuente: Elaboración propia.

Como se puede observar en las gráficas de medidas internas, el número óptimo de clústeres para el repositorio de archivos y la medida del coeficiente de *Silhouette* sugieren que son doce clústeres cuando se aplica la normalización y dos clústeres cuando no se aplica la normalización. Esto, debido a que el principio del coeficiente de *Silhouette* es: si el valor está más cercano a uno, el número de clúster es el más indicado; sin embargo, como se puede observar en las gráficas anteriores, los valores en el eje de las *y* son muy bajos, por lo que no es una buena métrica para tomar el número óptimo de clústeres, por lo tanto, se toma como referencia la medida interna de inercia para buscar los clústeres óptimos en el repositorio de archivos.

A continuación, se muestra las gráficas de clusterización de párrafos con K-Means aplicado al repositorio de archivos con y sin normalización mencionada anteriormente para comparar y ver las ventajas que puede proporcionar la normalización en análisis de texto o campos no estructurados.

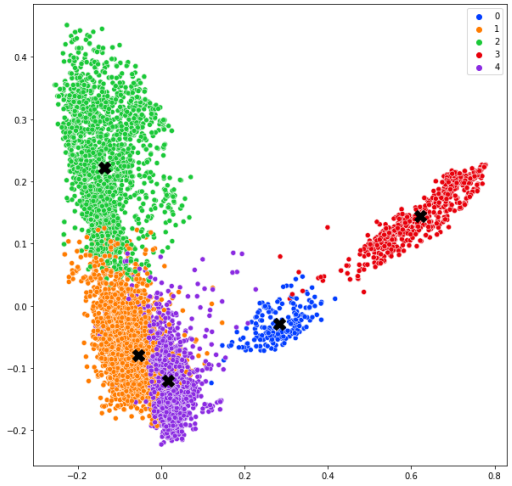


Figura 2.26: Agrupamiento con medida TF/IDF con normalización. Fuente: Elaboración propia.

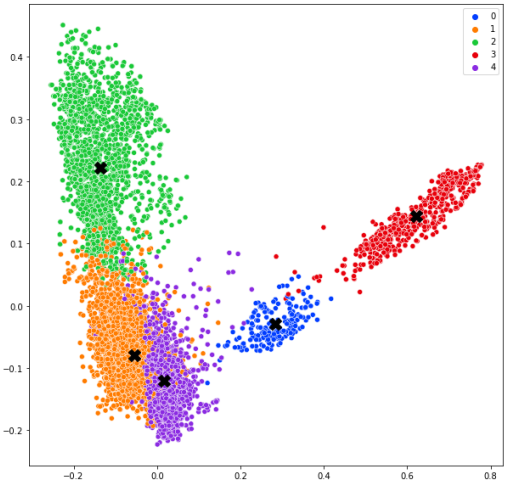


Figura 2.27: Agrupamiento con medida TF con normalización. Fuente: Elaboración propia.

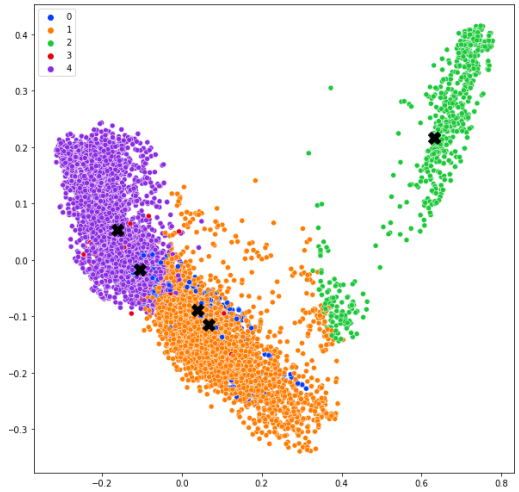


Figura 2.28: Agrupamiento con medida TF/IDF sin normalización. Fuente: Elaboración propia.

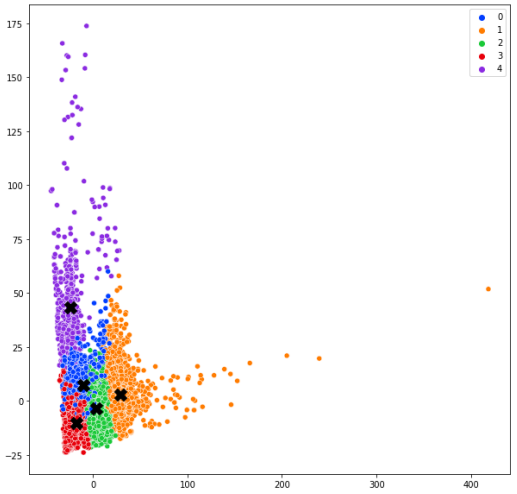


Figura 2.29: Agrupamiento con medida TF sin normalización. Fuente: Elaboración propia.

Al realizar las clusterización de párrafos con K-Means con seis y cinco clústeres se realiza mejor el agrupamiento con cinco que son los resultados mostrados en la graficas anteriores. La clusterización de párrafos con K-Means sin aplicar preprocesamiento

los clústeres que se definen no tienen una buena separación entre ellos y no se puede apreciar los cinco grupos establecidos en la medida como TF/IDF en comparación con los resultados obtenidos quitando los *stopwords*. y agregando la lematización.

2.2.4. Modelado de tópicos

Latent Dirichlet Allocation (LDA) es un algoritmo de aprendizaje no supervisado muy utilizado dentro del análisis de datos especialmente dentro del campo del PLN, por lo que adicional a la clusterización de párrafos con K-Means se aplica este algoritmo también dentro del análisis exploratorio y tener otra medida diferente de agrupación por medidas de similitud y comparar los resultados del modelado de tópicos con los resultados obtenidos con K-Means.

Para aplicar y modelar adecuadamente los subtemas en materia de PLD/FT, se usó el modelo LDA con diferentes números de tópicos en un rango de cuatro a diez. La mejor métrica se seleccionó utilizando tanto la medida de perplejidad, así como la proporción de palabras para cada tópico presentadas en los gráficos. Esta metodología permitió determinar con mayor precisión el número óptimo de temas en el repositorio de archivos oficiales de PLD/FT. A continuación, se muestran los resultados obtenidos.

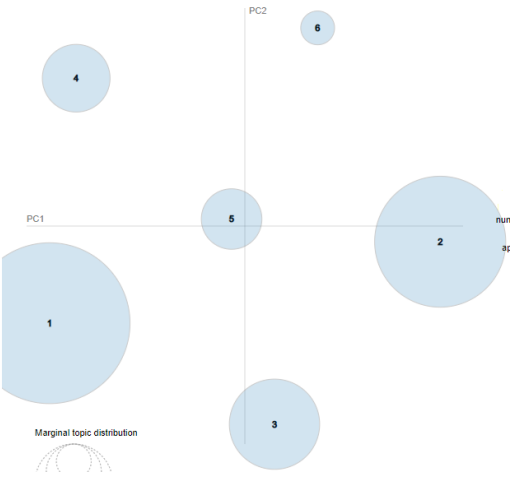


Figura 2.30: Modelo LDA con medida TF/IDF con normalización.
Fuente: Elaboración propia.

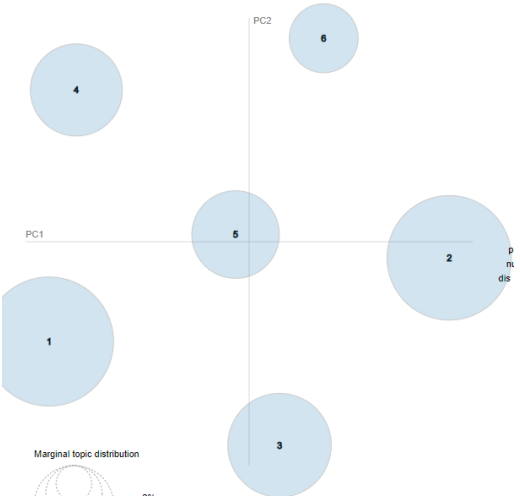


Figura 2.31: Modelo LDA con medida TF con normalización.
Fuente: Elaboración propia.

Figura 2.32: Modelo LDA con medida TF/IDF sin normalización
Fuente: Elaboración propia.

Figura 2.33: Modelo LDA con medida TF sin normalización.
Fuente: Elaboración propia.

Al utilizar el modelo LDA, no se observó una mejora significativa con respecto a la aplicación de normalización en comparación con el modelo K-Means. Sin embargo, se observó una mejora en la proporción de los temas al utilizar el algoritmo TF. El resultado del modelo de tópicos obtenido mediante LDA es similar al número de clústeres generados por el modelo K-Means, con la excepción de que en el modelo LDA los temas se separan completamente.

2.2.5. Análisis exploratorio de datos aplicado al conjunto de datos

El objetivo de aplicar análisis exploratorio sobre el conjunto de datos es obtener la estructura que tiene el mismo, por lo tanto, una vez definido el conjunto de datos se aplica este análisis para obtener estadística descriptiva sobre la información que serán los insumos para el entrenamiento y evaluación del modelo BERT que se implementa en el SBR *Especialista Virtual PLD*. A diferencia del análisis aplicado al repositorio de documentos, al conjunto de datos no se aplica ningún preprocesamiento, el EDA aplicado al conjunto de datos consiste en obtener longitudes y distribuciones de las siguientes propiedades del conjunto de datos:

1. Longitud de párrafos de todos los extraídos de los documentos del repositorio
2. Longitud de párrafos que se utilizaron para preguntas y respuesta.

3. Longitud de la pregunta a responder
4. Longitud de las respuestas a cada pregunta
5. Distribución del tipo de interrogativo en las preguntas
6. Distribución de prefijos de trigramas de preguntas

Cada uno de los histogramas se tomaron como referencia del análisis aplicado al conjunto de datos de SQuAD [45]. A continuación, se presentan los resultados obtenidos en histogramas de los seis datos estadísticos mencionados anteriormente.

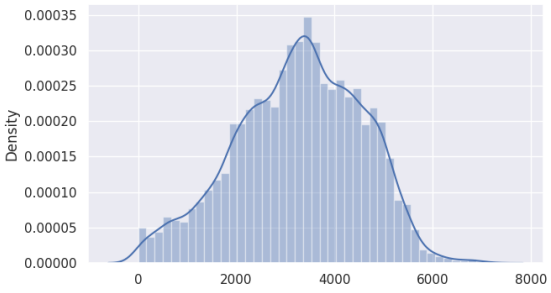


Figura 2.34: Longitud de párrafos en el repositorio de archivos.
Fuente: Elaboración propia.

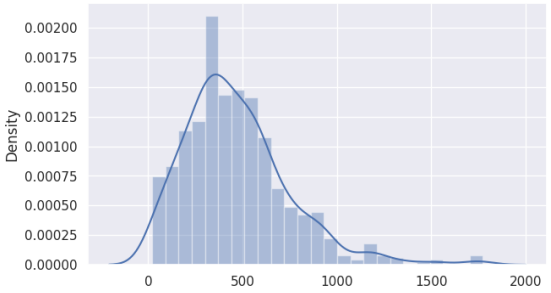


Figura 2.35: Longitud de párrafos en contexto de preguntas y respuestas.
Fuente: Elaboración propia.

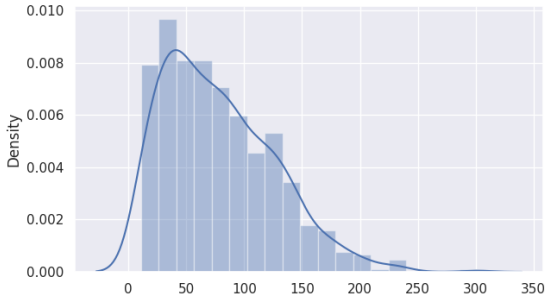


Figura 2.36: Longitud de preguntas en el conjunto de datos.
Fuente: Elaboración propia.

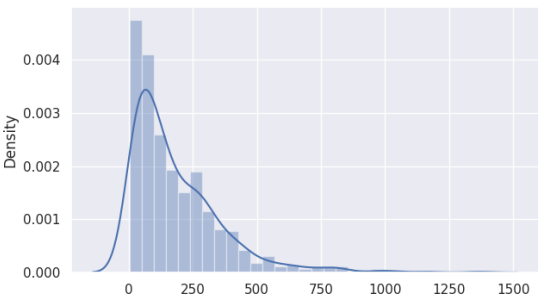


Figura 2.37: Longitud de respuestas en el conjunto de datos.
Fuente: Elaboración propia.

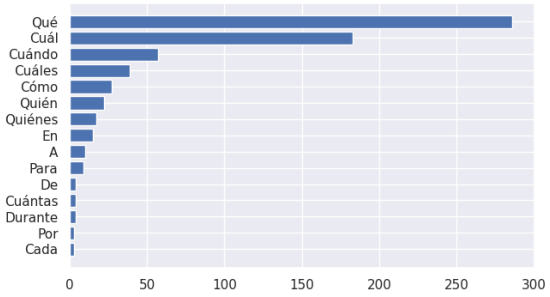


Figura 2.38: Distribución de preguntas.
Fuente: Elaboración propia.

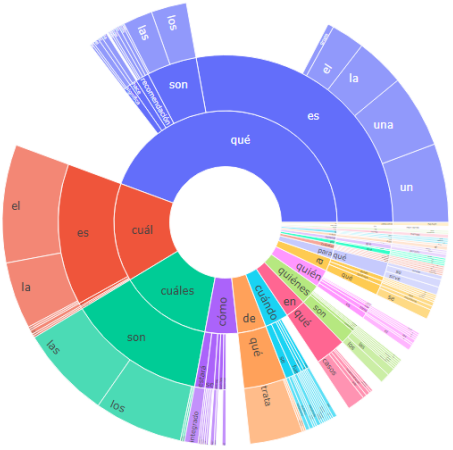


Figura 2.39: Distribución de prefijos de trigramas de preguntas.
Fuente: Elaboración propia.

En las gráficas anteriores se puede apreciar la distribución del conjunto de datos utilizado para entrenar y evaluar los modelos BERT. En las últimas dos se observa la distribución de preguntas y de prefijos de trigramas. Es evidente que existe una gran inclinación sobre las preguntas del tipo "¿Qué?" representan más de 250 preguntas. Dentro de estas, predominan las preguntas ¿Qué es? y ¿Qué son?

2.2.6. Conclusiones

El EDA aplicado al repositorio de archivos y conjunto de datos tenía un único propósito: analizar tanto al conjunto de datos como la documentación oficial relacionada con la regulación en materia de PLD/FT. Una vez completado este análisis, se concluye que, en el caso de los documentos dentro del repositorio de archivos, la clusterización funciona de mejor manera cuando se aplica la normalización de textos, lo cual se puede observar en las Figuras 2.26 y 2.27, donde los clústeres no se separan completamente, pero sí de manera más clara a cuando no se aplica la normalización. Además, el número de clústeres identificados en los documentos es similar al número de tópicos identificados mediante el algoritmo LDA, que en este caso identifica seis subtemas. Por tanto, podemos observar que, dentro de la documentación oficial, el algoritmo K-Means identifica cinco clústeres, mientras que LDA puede identificar seis subtemas en los documentos almacenados en el repositorio.

En cuanto al conjunto de datos, se generaron diversos histogramas para analizar cómo se distribuyen las preguntas. Se ha llegado a la conclusión de que las preguntas que comienzan con el interrogativo "¿Qué?"predominan en el conjunto de datos. Al revisar los histogramas del conjunto de datos, se observa que las respuestas tienen una longitud considerable, esto se da principalmente a la necesidad de definir conceptos. Esta característica podría dificultar la capacidad del modelo del *Especialista Virtual PLD* para responder correctamente. En cuanto a las preguntas planteadas en sí, se observa que contienen una cantidad considerable de texto y tienden a tener entre cincuenta y ciento veinte caracteres. Además, en lo que respecta a la longitud de los párrafos, se nota que estos son extensos, lo cual es una característica común en textos oficiales que abarcan diversos tipos de documentos.

Este análisis proporciona una visión general de cómo se estructuran los datos y las preguntas dentro del conjunto de información, lo que es esencial para la comprensión y el procesamiento efectivo de los mismos.

Capítulo 3

Descripción de modelo en materia de PLD/FT

Capítulo 3.

Sistemas de búsqueda de respuestas

3.1. Metodología para el sistema de búsqueda de respuestas

Los Sistemas de Preguntas y Respuestas (SBR) forman parte de un conjunto de tareas de PLN para el cual BERT proporciona modelos preentrenados que pueden ser personalizados y reentrenados en tópicos más especializados. El *Especialista Virtual PLD* utiliza dos de estos modelos al entrenar y evaluar un conjunto de datos en materia de Prevención de Lavado de Dinero y Financiamiento al Terrorismo (PLD/FT) en el idioma español.

El *Especialista Virtual PLD* propone principalmente la creación de un conjunto de datos en el idioma español en el dominio de PLD/FT. Para seleccionar el modelo que utiliza el *Especialista Virtual PLD*, la metodología plantea entrenar modelos BERT con dos variantes en específico, la primera con DistilBERT y la segunda con BETO, DistilBERT una versión más ligera y rápida que BERT conservando más del 95 % de rendimiento y BETO está pre entrenado en el idioma español. Ambos modelos BERT tendrán hiperparámetros estandarizados para el entrenamiento y evaluación del modelo para poder realizar una comparativa y seleccionar el mejor modelo.

La metodología para el entrenamiento y evaluación del modelo fue guiada por cuatro etapas, las cuales se describen a continuación:

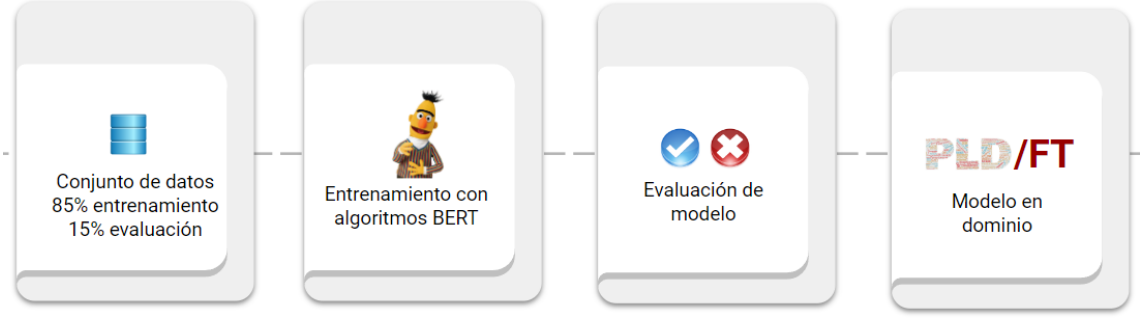


Figura 3.1: Metodología del Sistema de Búsqueda de Respuesta.
Fuente: Elaboración propia.

Etapa 1. Conjunto de datos para entrenamiento y evaluación con dominio en PLD/FT.

El primer paso de la metodología es la definición del conjunto de datos de datos el cual consta de 703 preguntas extraídas de documentos del repositorio de archivos PDF, las cuales, a su vez, fueron extraídas principalmente de documentos como: la Evaluación Nacional de Riesgos (ENR) del año 2020, las 40 Recomendaciones del GAFI y las Disposiciones de Carácter General para la regulación de PLD/FT.

Se toman principalmente de estos documentos por la información contenida en los mismos, como se mencionó en el capítulo 2, las Disposiciones de Carácter General (DCG) contienen todos los conocimientos técnicos en materia de PLD/FT; las 40 Recomendaciones definen un esquema completo para que los diferentes países puedan prevenir eficientemente el lavado de dinero; y, finalmente, la ENR proporciona los riesgos a los que se enfrenta México de LD/FT.

El conjunto de datos consta de preguntas donde la respuesta es el extracto de un párrafo dentro de un documento oficial y tiene diferentes tipos interrogativos, los principales están distribuidos de la siguiente manera:

Tipo de pregunta	Porcentaje de uso en preguntas
¿Qué?	41 %
¿Cuál?	26 %
¿Cuándo?	8 %
¿Cuáles?	6 %
¿Cómo?	4 %
¿Quién?	3 %
¿Quiénes?	2 %

Cuadro 3.1: Distribución de preguntas del conjunto de datos.
Fuente: Elaboración propia.

Para medir el desempeño de los modelos BERT para el SBR *Especialista Virtual PLD* el conjunto de datos se divide con una proporción de 85 % para el entrenamiento y 15 % para evaluación.

Etapa 2. Entrenamiento del modelo con algoritmos BERT.

Los algoritmos BERT basados en arquitectura de tipo *Transformer* han tomado mayor relevancia en los últimos años por los resultados obtenidos con grandes volúmenes de información. En la actualidad hay diferentes versiones y variaciones de modelos, la diferencia radica en el número de parámetros con los que fueron entrenados estos algoritmos, para este trabajo se seleccionaron dos, para las implementaciones de un SBR con dominio PLD/FT, los modelos seleccionados son los siguientes:

- DistilBERT
- BETO

Los entrenamientos con aprendizaje profundo (DL, por sus siglas en inglés) tienen la característica que entre más grande es el modelo, se obtienen mejores resultados, cada una de estas versiones seleccionadas tienen una arquitectura que permiten reducir el tamaño en el número de parámetros, aumentar la velocidad o están preentrenados en el idioma español, las primeras dos características conservando más del 95 % de rendimiento que proporciona la versión más completa de BERT (*bert-base-uncased*). La tercera característica es el modelo BETO preentrenado con el idioma español, tiene un mejor desempeño en comparación del multilenguaje de BERT. Para el entrenamiento y evaluación de un modelo de SBR, los conjuntos de datos generalmente se dividen en tres partes: entrenamiento, validación y pruebas, para este trabajo se define en solo dos partes: entrenamiento y pruebas, debido al tamaño del conjunto de datos, la proporción utilizada se define en la etapa 1.

Agregar vocabulario especializado en PLD/FT al tokenizador

Este trabajo presenta un conjunto de datos en materia de PLD/FT en el idioma español. En algunos casos, puede ser crucial enriquecer el vocabulario del modelo lingüístico y obtener palabras de dominios especializados como medicina, derecho, etc. Para lograr este enriquecimiento en el dominio especializado, se realizan ajustes al tokenizador de los modelos BERT, estos ajustes permiten la incorporación de vocabulario recurrente

relacionado con la temática sin modificar el vocabulario del tokenizador, de tal manera que se incorporan palabras completas descartando las subpalabras para finalmente enriquecer el vocabulario. Para realizar el aumento de vocabulario especializado, se realiza en un proceso de cinco pasos, como se muestra en la siguiente imagen:

Figura 3.2: Agregar vocabulario al tokenizador.
Fuente: Elaboración propia.

Paso 1. Extracción de información del repositorio de documentos oficiales en PLD/FT.

En este paso, se realiza un proceso de extracción de texto de todos los documentos almacenados en el repositorio de archivos. Los textos extraídos se guardan en un dataframe, donde cada párrafo se almacena como una unidad individual. Este dataframe se utiliza como insumo fundamental en el siguiente paso del proceso.

Es importante destacar que este proceso de extracción tiene similitudes con el utilizado en la clusterización de párrafos con el algoritmo de aprendizaje no supervisado K-Means. Ambos procesos se basan en la extracción de información textual para su posterior preprocesamiento y análisis. Al aplicar esta técnica de extracción en nuestro flujo de trabajo, nos aseguramos de tener acceso a los datos necesarios para realizar una exploración detallada de los documentos en el repositorio.

Paso 2. Preprocesamiento.

Para el paso dos, que implica la aplicación de preprocesamiento, se llevan a cabo una serie de tareas para preparar los datos de manera óptima. Primero, se eliminan las palabras vacías (*stopwords*), que son términos comunes que no aportan un significado sustancial al texto. Luego, se eliminan los signos de puntuación para mantener única-

mente las palabras relevantes. Además, se realiza un reemplazo de más de dos espacios en blanco por uno solo, esto permite a normalizar la estructura del texto, lo que facilita su procesamiento y aplicación de algoritmos de relevancia de términos que ayudan a conocer que tan relevante es una palabra para un documento en una colección.
Este proceso de preprocesamiento es esencial, ya que nos permite obtener un vocabulario exclusivo y relevante relacionado con la temática de PLD/FT. Al eliminar las palabras vacías y los signos de puntuación, y al normalizar la estructura del texto, obtenemos un conjunto de palabras clave en este dominio.

Paso 3. Vectorización de palabras.
El tercer paso implica la vectorización de palabras que es el insumo para la aplicación del algoritmo TF/IDF, por lo que se utilizó un tokenizador, que divide el texto en unidades más pequeñas, en este caso palabras, y las asigna a un índice único. Esto nos permite representar cada palabra como un vector numérico en una matriz y construir una matriz de palabras a partir del vocabulario obtenido único extraído de los documentos oficiales dentro de la regulación PLD/FT.

Paso 4. Aplicación del algoritmo TF/IDF.
Con la matriz ya establecida, se aplica el algoritmo de frecuencias de términos TF/IDF para identificar las palabras más relevantes en el corpus PLD/FT. Este algoritmo asigna una puntuación a cada palabra en función de su frecuencia en un documento específico (TF) y de su relevancia en todo el corpus (IDF). Las palabras con puntuaciones más altas se consideraron más relevantes para el tema de PLD/FT.
Estas palabras clave identificadas mediante TF/IDF se agregan al tokenizador, enriqueciendo así el vocabulario utilizado para la representación numérica de los documentos. Al agregar estas palabras clave, el tokenizador adquiere una mayor capacidad para capturar términos significativos relacionados con PLD/FT durante el entrenamiento de los modelos DistilBERT y BETO.

Paso 5. Redimensionamiento de Embeddings.

Finalmente, se procede a redimensionar el tamaño de la matriz de *embeddings* con los nuevos términos relevantes en materia de PLD/FT agregados al vocabulario. Una vez redimensionada la matriz, se llevan a cabo nuevamente los experimentos establecidos, pero ahora teniendo en cuenta este cambio en el tokenizador de BERT.

Con este ajuste en el tokenizador de BERT, que implica la incorporación de los nuevos términos de PLD/FT al vocabulario y la redimensión de la matriz de *embeddings*, se espera mejorar la métrica de los resultados en cada uno de nuestros experimentos.

Este cambio permitirá que nuestro modelo procese de manera más efectiva los términos relacionados con PLD/FT, lo que debería reflejarse en un aumento en la calidad de los resultados obtenidos.

Etapa 3. Evaluación del modelo

El tercer paso de esta metodología es la evaluación del modelo de Sistema Búsqueda de Respuestas (SBR), que consiste en evaluar la capacidad de aprendizaje del modelo BERT de una forma cuantitativa. Para esto, se utilizan las métricas de: *Exact Match* y *F1* para medir el desempeño de los modelos BERT y tomar el tenga mejor desempeño.

Con estas métricas de precisión, se selecciona el modelo que presenta el mejor rendimiento al evaluar el conjunto de datos en materia de PLD/FT. Para cada entrenamiento del modelo, se establece una configuración de hiperparámetros estandarizada y se realizan experimentos con diferentes valores para obtener los mejores resultados en la evaluación del conjunto de datos en materia de PLD/FT. Los hiperparámetros elegidos son los siguientes:

1. Número de épocas (*epoch*) {20, 30, 40, 50}
2. Tamaño por lotes (*batch size*) {16, 32, 48, 64}
3. Tasa de aprendizaje (*learning rate*) {5e-5, 3e-5, 2e-5}
4. Máximo largo de secuencia (*max length*) {512, 384}
5. Dilución (*dropout*) {0.1, 0.3}
6. Función de activación (*function activation*) {gelu, relu, silu}

En cada uno de los experimentos llevados a cabo con los modelos definidos y los hiperparámetros establecidos, se comparan las métricas de *F1* y *Exact Match*. De esta manera, se selecciona el modelo que mejor desempeño haya mostrado para el SBR en el dominio de PLD/FT.

Etapa 4. Modelo con dominio PLD/FT

Finalmente, el último paso de la metodología consiste en definir un modelo con dominio PLD/FT utilizando la versión del algoritmo BERT que ha demostrado un mejor rendimiento durante su entrenamiento y evaluación con las métricas de precisión *F1* y *Exact Match*. Esta elección se basa en la comparación de modelos y diferentes hiperparámetros realizada en la etapa anterior. En la etapa final del proyecto, se pone a disposición el modelo entrenado en materia de PLD/FT desde aspectos básico, datos relevantes y conocimiento técnicos en esta regulación.

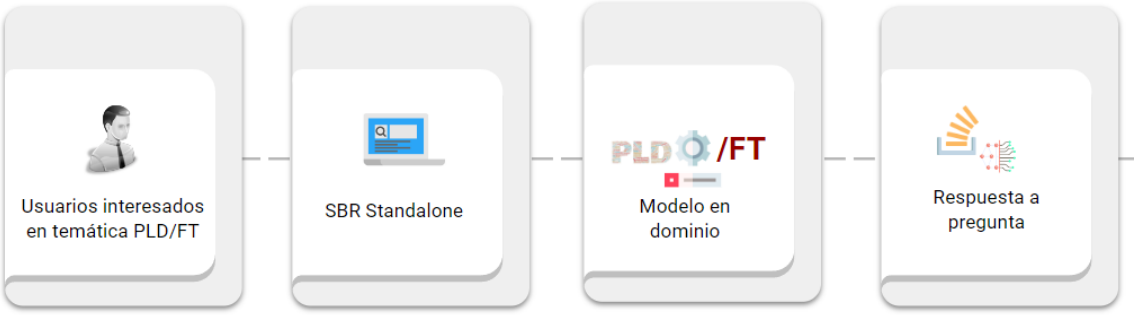


Figura 3.3: Flujo de aplicación Especialista Virtual PLD.
Fuente: Elaboración propia.

Capítulo 4

Resultados y análisis

Capítulo 4.

Resultados y análisis

4.1. Evaluación de modelos BERT

En este capítulo, se presentan los resultados obtenidos en la evaluación del rendimiento de los modelos BERT: DistilBERT [39] y BETO [9], en el contexto específico de la Prevención de Lavado de Dinero y Financiamiento al Terrorismo (PLD/FT), objetivo principal de este trabajo. En el marco de esta investigación, se explora el rendimiento de los modelos mencionados al evaluar un conjunto de datos específico en materia de PLD/FT. Con este propósito, se sigue una metodología que incorpora el enfoque propuesto por *Hugging Face*, haciendo uso del entrenamiento de un modelo para un conjunto de datos personalizado conocido como "*Fine-tuning with custom datasets*"[4]; esta técnica posibilita el entrenamiento y la evaluación de un modelo optimizado para la tarea un sistemas de búsqueda y respuestas que se incorpora al *Especialista Virtual en PLD*.

Como parte de los resultados de este trabajo, se toma como referencia SQUAD [36] para la evaluación de un conjunto de datos en materia de PLD/FT, debido a la relevancia y éxito que ha demostrado SQUAD en la comunidad de procesamiento de lenguaje natural, especialmente en la evaluación de modelos de sistemas de búsqueda de respuestas. Además, es importante destacar que SQUAD realizó una implementación con la arquitectura de DistilBERT [39], uno de los modelos seleccionados para la evaluación del conjunto de datos en cuestión. Para llevar a cabo esta evaluación, se hace uso del código publicado por SQUAD con DistilBERT [5] para evaluar un conjunto de datos específico en materia de la PLD/FT.

Es fundamental destacar que todas las evaluaciones y desarrollos se implementaron en un entorno tecnológico optimizado, basado en Python 3.9 y aprovechando la potencia

[4] Hugging Face. *Fine-tuning with custom datasets*. (s.f.). https://Huggingface.co/transformers/v3.2.0/custom_datasets.html

[5] GitHub. Question-Answering-SQUAD. (s.f.). https://github.com/nlpunibo/Question-Answering-SQUAD/blob/main/DistilbertQA_train.ipynb

computacional de una GPU T4, en el entorno de Google Colab. Este entorno proporciona las herramientas y los recursos necesarios para el entrenamiento y evaluación de los modelos en cuestión.

En este capítulo, se listan los detalles de los resultados, análisis y conclusiones de la evaluación de los modelos. El propósito es mostrar la eficacia de los modelos BERT en el dominio específico en PLD/FT con un conjunto de datos en este campo especializado.

El entrenamiento y la evaluación de cada modelo se llevaron a cabo utilizando una combinación de hiperparámetros estandarizados, con el fin de permitir una comparación significativa entre ellos. Los resultados del rendimiento obtenidos en cada modelo se presentan de la siguiente manera:

1. Evaluación de los modelos DistilBERT y BETO utilizando las versiones *base uncased*.
2. Evaluación de los modelos BERT: DistilBERT y BETO, utilizando las versiones *base uncased* con la incorporación de palabras relevantes en materia de PLD/FT al tokenizador de BERT.
3. Evaluación del conjunto de datos utilizando el modelo definido por SQUAD y la combinación de hiperparámetros establecidos por dicho proyecto.

Los resultados de estos experimentos se presentan en tablas que muestran la precisión con las medidas de evaluación *F1* y *Exact Match*. El capítulo 3 proporciona una descripción detallada de los valores específicos asignados a cada hiperparámetro en cada uno de los experimentos. Esto incluye los siguientes la *activation function*, *learning rate*, *batch size*, *max length* y *epoch*. Además de estos, se utilizó el algoritmo de optimización *Adam* para cada uno de los experimentos realizados.

4.1.1. DistilBERT y BETO en su versión uncased

4.1.1.1. DistilBERT

Considerando los modelos específicos de BERT mencionados, se implementó una estrategia de conversión en minúsculas al conjunto de datos con el propósito de emplear

la versión *base uncased* en ambos modelos. Los experimentos que se describirán se centran en el modelo DistilBERT, en el cual se exploraron diversas combinaciones de hiperparámetros. Este enfoque buscó evaluar la precisión del modelo, con el objetivo fundamental consolidar los resultados. Con la comparación de los resultados obtenidos por el modelo DistilBERT en cada etapa de entrenamiento, se logró una evaluación completa de su rendimiento.

En esta sección se describen los resultados obtenidos en las diversas combinaciones de hiperparámetros evaluados para el modelo DistilBERT, y la comparación detallada de su rendimiento. Esta exploración permitió identificar la mejor combinación de hiperparámetros para posteriormente compararlos con el modelo BETO, estos hiperparámetros ayudan a facilitar la elección del modelo más adecuado para *Especialista Virtual PLD*. Basándonos en los experimentos realizados, la tabla 4.1 resume los mejores resultados obtenidos con el modelo DistilBERT. Los resultados obtenidos utilizando diferentes hiperparámetros son muy similares. La precisión más baja registrada para la métrica de *Exact Match* fue de 25.47 %, mientras que para la métrica *F1* fue de 32.07 %. A modo de resumen, se presentan los resultados más destacados obtenidos por época al comparar todas las combinaciones de hiperparámetros en cada experimento, se tomó como mejores resultados la precisión obtenida en la métrica de *Exact Match*. Se hace especial énfasis en la función de pérdida y las puntuaciones en las métricas de *Exact Match* y *F1*. Dentro de estas diversas combinaciones, se observa que el rendimiento óptimo se alcanza con 50 épocas, utilizando los siguientes hiperparámetros: *learning rate* de 2e-5, *max length* de 512, *batch size* de 16, *dropout* de 0.3 y *activation function gelu*.

Hiperparámetros	Función de pérdida	Exact Match	F1	Época
{learning rate: 5e-5, max length: 384, batch size: 32, dropout: 0.3, activation function: gelu}	0.0513	0.36793	0.5578	20
{learning rate: 2e-5, max length: 512, batch size: 32, dropout: 0.1, activation function: gelu}	0.000089	0.386792	0.523585	30
{learning rate: 2e-5, max length: 512, batch size: 32, dropout: 0.3, activation function: gelu}	0.0002	0.3679	0.5425	40
{**learning rate: 2e-5, max length: 512, batch size: 16, dropout: 0.3, activation function: gelu**}	**0.000134**	**0.386792**	**0.54717**	**50**

Cuadro 4.1: Resultados obtenidos con modelo DistilBERT.
Fuente: Elaboración propia.

La finalidad principal de llevar a cabo los entrenamientos utilizando el modelo DistilBERT fue obtener los resultados óptimos para la posterior comparación con los resultados obtenidos con el modelo BETO. Con este propósito en mente, se ejecutaron un total de 96 experimentos, abarcando diversas combinaciones de hiperparámetros. Este proceso de experimentación buscó identificar las configuraciones que proporcionaran el mejor rendimiento. Sin embargo, al examinar los resultados presentados en el cuadro 4.1, se hace evidente que las métricas *Exact Match* y *F1* no arrojaron resultados favorables. En específico, se observa que tanto la precisión *Exact Match* como la puntuación *F1* no lograron alcanzar niveles satisfactorios. La métrica *Exact Match* reportó un máximo de 38.67 % de precisión, mientras que la puntuación *F1* se situó en 54.71 % dentro de la misma ejecución, tomando el valor máximo en dicha métrica alcanzó un valor máximo de 56.04 % pero con 36.79 % en Exact Match. Ambos resultados, lamentablemente, presentaron valores bajos para un sistema de preguntas y respuestas.
A continuación, se enumeran todos los experimentos realizados con diferentes hiperparámetros y la versión *base uncased* para el modelo DistilBERT.

Resultados de experimentos con el modelo DistilBERT en 20 épocas

A continuación, se presentan las siguientes tablas con los resultados obtenidos mediante el modelo DistilBERT. Estas tablas se enfocan en los experimentos realizados en 20 épocas, utilizando las combinaciones de hiperparámetros establecidas en el capítulo 3. Con el propósito de mejorar la presentación de los resultados en dichas tablas, se han abreviado los hiperparámetros *batch size* y *max length* a dos caracteres, denominándolos como *bs* y *ml*, respectivamente.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.330189	0.273585	0.320755	0.495283	0.476415	0.481132	0.358491	0.320755	0.320755	0.533019	0.485849	0.476415
384	32	0.339623	0.311321	0.320755	0.523585	0.462264	0.485849	**0.367925**	0.311321	0.320755	**0.557812**	0.481132	0.481132
384	48	0.301887	0.320755	0.311321	0.448113	0.485849	0.471698	0.320755	0.367925	0.311321	0.466981	0.504717	0.509434
384	64	0.301887	0.330189	0.330189	0.490566	0.471698	0.481132	0.330189	0.320755	0.330189	0.509434	0.481132	0.514151
512	16	0.358491	0.273585	0.320755	0.509434	0.448113	0.466981	0.320755	0.311321	0.320755	0.320755	0.466981	0.481132
512	32	0.320755	0.301887	0.301887	0.481132	0.476415	0.476415	0.330189	0.301887	0.301887	0.476415	0.45283	0.485849
512	48	0.320755	0.283019	0.292453	0.490566	0.448113	0.443396	0.330189	0.358491	0.292453	0.481132	0.509434	0.462264
512	64	0.301887	0.320755	0.292453	0.471698	0.471698	0.45283	0.358491	0.320755	0.292453	0.5	0.481132	0.476415

Cuadro 4.2: DistilBERT y learning rate a 5e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.349057	0.283019	0.292453	0.514151	0.443396	0.443396	**0.358491**	0.301887	0.292453	**0.537736**	0.481132	0.466981
384	32	0.339623	0.320755	0.320755	0.509434	0.485849	0.45283	0.339623	0.311321	0.320755	0.509434	0.501562	0.509434
384	48	0.301887	0.320755	0.330189	0.471698	0.495283	0.471698	0.358491	0.339623	0.330189	0.509434	0.476415	0.514151
384	64	0.320755	0.320755	0.311321	0.485849	0.476415	0.457547	0.349057	0.330189	0.311321	0.509434	0.490566	0.485849
512	16	0.349057	0.320755	0.301887	0.485849	0.471698	0.5	0.330189	0.320755	0.301887	0.504717	0.471698	0.504717
512	32	0.330189	0.330189	0.330189	0.495283	0.45283	0.457547	0.349057	0.349057	0.330189	0.5375	0.5	0.509434
512	48	0.292453	0.264151	0.301887	0.481132	0.45283	0.5	0.330189	0.320755	0.301887	0.471698	0.490566	0.471698
512	64	0.339623	0.254717	0.339623	0.504717	0.45283	0.466981	0.330189	0.330189	0.339623	0.504717	0.495283	0.504717

Cuadro 4.3: DistilBERT y learning rate a 3e-5
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.320755	0.283019	0.330189	0.5	0.438679	0.466981	0.301887	0.301887	0.330189	0.5	0.462264	0.45283
384	32	0.320755	0.330189	0.320755	0.471698	0.476415	0.471698	0.339623	0.349057	0.320755	0.481132	**0.517969**	0.471698
384	48	0.320755	0.339623	0.292453	0.476415	0.462264	0.433962	0.301887	0.264151	0.292453	0.471698	0.45283	0.504717
384	64	0.330189	0.320755	0.311321	0.490566	0.457547	0.45283	0.330189	0.311321	0.311321	0.495283	0.457547	0.509434
512	16	0.339623	0.301887	0.320755	0.490566	0.45283	0.466981	0.349057	0.301887	0.320755	0.5	0.462264	0.457547
512	32	0.339623	0.301887	0.339623	0.476415	0.448113	0.481132	0.330189	0.339623	0.339623	0.481132	0.490566	0.462264
512	48	0.311321	0.283019	0.320755	0.466981	0.438679	0.429245	0.311321	0.273585	0.320755	0.471698	0.466981	0.462264
512	64	0.330189	0.320755	0.330189	0.481132	0.462264	0.45283	**0.358491**	0.320755	0.330189	0.495283	0.490566	0.462264

Cuadro 4.4: DistilBERT y learning rate a 2e-5.
Fuente: Elaboración propia.

Resultados de experimentos con el modelo DistilBERT en 30 épocas

A continuación, se presentan las siguientes tablas con los resultados obtenidos mediante el modelo DistilBERT. Estas tablas se enfocan en los experimentos realizados en 30 épocas, utilizando las combinaciones de hiperparámetros establecidas en el capítulo 3. Con el propósito de mejorar la presentación de los resultados en dichas tablas, se han abreviado los hiperparámetros *batch size* y *max length* a dos caracteres, denominándolos como *bs* y *ml*, respectivamente.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.358491	0.301887	0.320755	0.528302	0.462264	0.481132	0.358491	0.358491	0.320755	0.523585	0.514151	0.485849
384	32	0.339623	0.330189	0.339623	0.528302	0.471698	0.481132	0.292453	0.339623	0.339623	0.495283	0.514151	0.504717
384	48	0.301887	0.311321	0.311321	0.525694	0.462264	0.498611	0.311321	0.358491	0.311321	0.485849	0.509434	0.476415
384	64	0.330189	0.320755	0.301887	0.498512	0.490566	0.462264	0.320755	0.358491	0.301887	0.504717	0.495283	0.462264
512	16	0.349057	0.292453	0.320755	0.504717	0.481132	0.481132	0.349057	0.339623	0.320755	0.537736	0.490566	0.509434
512	32	0.320755	0.320755	0.320755	0.514151	0.476415	0.481132	**0.367925**	0.301887	0.320755	**0.542453**	0.476415	0.504717
512	48	0.339623	0.292453	0.292453	0.504717	0.45283	0.462264	0.330189	0.330189	0.292453	0.485849	0.490566	0.509434
512	64	0.320755	0.283019	0.311321	0.466981	0.476415	0.481132	0.358491	0.339623	0.311321	0.542453	0.5	0.504717

Cuadro 4.5: DistilBERT y learning rate a 5e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.339623	0.339623	0.339623	0.504717	0.490566	0.5	0.349057	0.339623	0.339623	0.532143	0.5	0.490566
384	32	0.339623	0.301887	0.349057	0.528302	0.462264	0.490566	0.330189	0.339623	0.349057	0.490566	0.509434	0.490566
384	48	0.283019	0.339623	0.320755	0.466981	0.481132	0.485849	0.320755	0.330189	0.320755	0.490566	0.5	0.504717
384	64	0.311321	0.311321	0.330189	0.481132	0.462264	0.476415	0.320755	0.339623	0.330189	0.509434	0.495283	0.466981
512	16	0.311321	**0.386792**	0.330189	0.496429	0.523585	0.471698	0.358491	0.311321	0.330189	**0.536607**	0.490566	0.466981
512	32	0.320755	0.367925	0.301887	0.505469	0.481132	0.471698	0.330189	0.320755	0.301887	0.523585	0.476415	0.490566
512	48	0.330189	0.292453	0.311321	0.495283	0.457547	0.457547	0.339623	0.330189	0.311321	0.523585	0.490566	0.485849
512	64	0.320755	0.283019	0.292453	0.466981	0.495283	0.481132	0.349057	0.330189	0.292453	0.523438	0.495283	0.495283

Cuadro 4.6: DistilBERT y learning rate a 3e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.339623	0.292453	0.330189	0.518868	0.476415	0.471698	0.367925	0.358491	0.330189	0.504717	0.495283	0.514151
384	32	0.349057	0.311321	0.311321	**0.545313**	0.471698	0.476415	0.349057	0.339623	0.311321	0.523585	0.495283	0.523585
384	48	0.311321	0.320755	0.320755	0.490566	0.485849	0.481132	0.330189	0.339623	0.320755	0.490566	0.5	0.509434
384	64	0.330189	0.301887	0.320755	0.485849	0.485849	0.476415	0.330189	0.358491	0.320755	0.481132	0.509434	0.528302
512	16	0.349057	0.320755	0.311321	0.5	0.495283	0.481132	0.330189	0.330189	0.311321	0.504717	0.481132	0.5
512	32	**0.386792**	0.330189	0.292453	0.523585	0.481132	0.471698	0.339623	0.349057	0.292453	0.523585	0.509434	0.504717
512	48	0.339623	0.30188	0.320755	0.509434	0.457547	0.481132	0.349057	0.311321	0.320755	0.490566	0.5125	0.490566
512	64	0.330189	0.311321	0.311321	0.5	0.5	0.462264	0.311321	0.339623	0.311321	0.481132	0.509434	0.504717

Cuadro 4.7: DistilBERT y learning rate a 2e-5.
Fuente: Elaboración propia.

Resultados de experimentos con el modelo DistilBERT en 40 épocas

A continuación, se presentan las siguientes tablas con los resultados obtenidos mediante el modelo DistilBERT. Estas tablas se enfocan en los experimentos realizados en 40 épocas, utilizando las combinaciones de hiperparámetros establecidas en el capítulo 3. Con el propósito de mejorar la presentación de los resultados en dichas tablas, se han abreviado los hiperparámetros *batch size* y *max length* a dos caracteres, denominándolos como *bs* y *ml*, respectivamente.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.339623	0.301887	0.283019	0.528302	0.448113	0.429245	0.339623	0.330189	0.283019	0.495283	0.504717	0.490566
384	32	0.339623	0.301887	0.292453	0.514151	0.457547	0.471698	0.311321	0.339623	0.292453	0.485849	0.490566	0.495283
384	48	0.339623	0.339623	0.292453	0.509434	0.485849	0.471698	0.349057	0.349057	0.292453	0.523585	0.5	0.523585
384	64	0.330189	0.349057	0.273585	0.495283	0.490566	0.448113	0.320755	0.330189	0.273585	0.490327	0.502232	0.495283
512	16	0.330189	0.311321	0.330189	0.481132	0.476415	0.504717	0.339623	0.311321	0.330189	0.514151	0.481132	0.485849
512	32	0.349057	0.283019	0.311321	0.5	0.466981	0.471698	0.349057	0.320755	0.311321	**0.528302**	0.495283	0.518868
512	48	0.320755	0.311321	0.301887	0.5	0.481132	0.45283	0.330189	0.320755	0.301887	0.490566	0.481132	0.509434
512	64	**0.367925**	0.301887	0.292453	0.502232	0.476415	0.438679	0.320755	0.311321	0.292453	0.504717	0.485849	0.471698

Cuadro 4.8: DistilBERT y learning rate a 5e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.349057	0.301887	0.283019	0.528302	0.45283	0.462264	**0.367925**	0.339623	0.283019	0.5	0.490566	0.495283
384	32	0.320755	0.320755	0.292453	0.504717	0.466981	0.462264	0.311321	0.339623	0.292453	0.476415	0.5	0.504717
384	48	0.311321	0.339623	0.311321	0.495283	0.490566	0.462264	0.339623	0.339623	0.311321	0.485849	0.495283	0.5
384	64	0.339623	0.301887	0.330189	0.485849	0.504717	0.481132	0.349057	0.349057	0.330189	0.504717	0.516183	0.5
512	16	0.339623	0.330189	0.320755	0.485849	0.485849	0.495283	0.358491	0.349057	0.320755	0.518868	0.5	0.476415
512	32	0.339623	0.311321	0.349057	0.495283	0.466981	0.481132	0.349057	0.349057	0.349057	0.528302	0.504717	0.518868
512	48	0.349057	0.301887	0.320755	0.485849	0.462264	0.481132	0.339623	0.339623	0.320755	0.509434	0.522917	0.495283
512	64	0.320755	0.311321	0.311321	0.481132	0.443396	0.5	0.349057	0.339623	0.311321	**0.537736**	0.495283	0.504717

Cuadro 4.9: DistilBERT y learning rate a 3e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.367925	0.330189	0.330189	0.533019	0.471698	0.485849	0.330189	0.320755	0.330189	0.542453	0.481132	0.5
384	32	0.330189	0.330189	0.349057	0.514151	0.485849	0.5	0.349057	0.330189	0.349057	0.542453	0.514063	0.5
384	48	0.311321	0.311321	0.349057	0.490566	0.481132	0.504717	0.320755	0.339623	0.349057	0.533019	0.504717	0.5
384	64	0.330189	0.311321	0.330189	0.485849	0.478423	0.485849	0.339623	0.339623	0.330189	0.528302	0.5	0.485849
512	16	0.330189	0.301887	0.330189	0.504717	0.443396	0.5	0.349057	0.330189	0.330189	0.542453	0.509434	0.495283
512	32	0.301887	0.320755	0.311321	0.495283	0.471698	0.481132	**0.367925**	0.349057	0.311321	**0.542453**	0.514151	0.495283
512	48	0.358491	0.292453	0.301887	0.514151	0.476415	0.471698	0.339623	0.292453	0.301887	0.533019	0.476415	0.481132
512	64	0.301887	0.283019	0.320755	0.476415	0.476415	0.485849	0.358491	0.339623	0.320755	0.528302	0.495283	0.509434

Cuadro 4.10: DistilBERT y learning rate a 2e-5.
Fuente: Elaboración propia.

Resultados de experimentos con el modelo DistilBERT en 50 épocas

A continuación, se presentan las siguientes tablas con los resultados obtenidos mediante el modelo DistilBERT. Estas tablas se enfocan en los experimentos realizados en 50 épocas, utilizando las combinaciones de hiperparámetros establecidas en el capítulo 3. Con el propósito de mejorar la presentación de los resultados en dichas tablas, se han abreviado los hiperparámetros *batch size* y *max length* a dos caracteres, denominándolos como *bs* y *ml*, respectivamente.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.320755	0.292453	0.301887	0.490566	0.466981	0.490566	0.330189	0.301887	0.301887	0.509434	0.476415	0.476415
384	32	0.349057	0.330189	0.273585	0.5	0.485849	0.443396	0.330189	0.339623	0.273585	0.495283	0.490566	0.476415
384	48	0.301887	0.311321	0.339623	0.476415	0.471698	0.5	0.339623	0.339623	0.339623	0.504717	0.481132	0.485849
384	64	0.311321	0.311321	0.320755	0.476415	0.495283	0.476415	0.358491	0.330189	0.320755	0.523585	0.504717	0.481132
512	16	0.330189	0.311321	0.339623	0.490566	0.471698	0.481132	0.320755	0.301887	0.339623	0.5	0.471698	0.490566
512	32	0.330189	0.320755	0.339623	0.495283	0.490566	0.481132	0.377358	0.349057	0.339623	0.528302	0.509434	0.495283
512	48	0.358491	0.292453	0.311321	**0.539583**	0.476415	0.476415	0.320755	0.330189	0.311321	0.320755	0.5	0.485849
512	64	**0.386792**	0.301887	0.320755	0.523585	0.462264	0.457547	0.320755	0.358491	0.320755	0.476415	0.523996	0.5

Cuadro 4.11: DistilBERT y learning rate a 5e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.349057	0.320755	0.330189	0.504717	0.5	0.485849	0.311321	0.311321	0.330189	0.485849	0.476415	0.490566
384	32	0.358491	0.301887	0.283019	0.514151	0.462264	0.448113	0.320755	0.330189	0.283019	0.504717	0.476415	0.490566
384	48	0.330189	0.330189	0.301887	0.471698	0.481132	0.476415	**0.367925**	0.330189	0.301887	**0.560417**	0.509434	0.485849
384	64	0.339623	0.320755	0.330189	0.518868	0.518868	0.495283	0.349057	0.330189	0.330189	0.502418	0.490566	0.481132
512	16	0.330189	0.320755	0.339623	0.495283	0.514151	0.457547	0.330189	0.339623	0.339623	0.528302	0.490566	0.471698
512	32	0.311321	0.349057	0.330189	0.504717	0.457547	0.476415	0.349057	0.301887	0.330189	0.533019	0.462264	0.485849
512	48	0.339623	0.301887	0.311321	0.529861	0.476415	0.466981	0.339623	0.320755	0.311321	0.533019	0.495283	0.495283
512	64	0.349057	0.311321	0.301887	0.509434	0.490566	0.481132	0.339623	0.330189	0.301887	0.537736	0.495283	0.495283

Cuadro 4.12: DistilBERT y learning rate a 3e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.339623	0.301887	0.311321	0.514151	0.457547	0.45283	0.311321	0.311321	0.311321	0.54717	0.481132	0.504717
384	32	0.330189	0.292453	0.311321	0.514151	0.457547	0.462264	0.330189	0.330189	0.311321	0.523585	0.495283	0.521875
384	48	0.301887	0.311321	0.292453	0.476415	0.504717	0.466981	0.320755	0.349057	0.292453	0.514151	0.509434	0.448113
384	64	0.320755	0.320755	0.330189	0.490566	0.476415	0.485849	0.349057	0.339623	0.330189	0.511719	0.508185	0.466981
512	16	0.320755	0.330189	0.330189	0.490566	0.490566	0.481132	**0.386792**	0.320755	0.330189	**0.54717**	0.495283	0.490566
512	32	0.349057	0.330189	0.339623	0.495283	0.476415	0.509434	0.349057	0.330189	0.339623	0.523585	0.490566	0.490566
512	48	0.339623	0.292453	0.292453	0.504717	0.462264	0.438679	0.330189	0.311321	0.292453	0.514151	0.495283	0.495283
512	64	0.301887	0.292453	0.292453	0.476415	0.476563	0.45283	0.330189	0.311321	0.292453	0.511719	0.495283	0.476415

Cuadro 4.13: DistilBERT y learning rate a 2e-5.
Fuente: Elaboración propia.

4.1.1.2. BETO

Continuado con la metodología establecida para la evaluación del conjunto de datos en materia de PLD/FT, BETO es otro modelo BERT seleccionado para las pruebas dentro del marco del Sistema de Preguntas y Respuestas (SBR) del *Especialista Virtual PLD*. La inclusión de BETO en esta evaluación se detalla en la segunda etapa de la metodología de este proyecto, específicamente en el apartado titulado "Entrenamiento del modelo con algoritmos BERT". Un aspecto relevante de este modelo es que fue preentrenado en el idioma español, lo que añade una característica especialmente importante a su evaluación en un contexto específico como el de PLD/FT.

El propósito de estas pruebas radica en la búsqueda de la configuración que maximice el rendimiento del modelo dentro del contexto del SBR en materia de PLD/FT, por lo tanto, de manera análoga al enfoque seguido con DistilBERT, se llevaron a cabo los mismos experimentos con BETO, los cuales involucraron la exploración de diversas configuraciones de hiperparámetros e iteraciones en el número de épocas de entrenamiento. Los resultados obtenidos en estos experimentos se enfocó en métricas clave como *Exact Match* y *F1*.

Con el objetivo de comparar los resultados obtenidos con BETO respecto a DistilBERT, se comparan los 96 experimentos realizados. Esta metodología tuvo como finalidad identificar las configuraciones de hiperparámetros que ofrecieran el mejor rendimiento en ambos modelos y así permitir una evaluación equitativa de sus capacidades.

A modo de resumen, se presentan los mejores resultados que se obtuvieron por época al comparar todas las combinaciones de hiperparámetros con el modelo BETO. Se hace especial énfasis en la función de pérdida y las puntuaciones en las métricas de *Exact Match* y *F1*. Dentro de estas diversas combinaciones, se observa que el mejor rendimiento se alcanza con 50 épocas, utilizando los siguientes hiperparámetros: *learning rate* de 5e-5, *max length* de 512, *batch size* de 32, *dropout* de 0.1 y *activation function gelu*.

Hiperparámetros	Función de pérdida	Exact Match	F1	Época
{learning rate: 3e-5, max length: 384, batch size: 48, dropout: 0.1, activation function: relu}	0.000002	0.386792	0.518868	20
{learning rate: 3e-5, max length: 384, batch size: 64, dropout: 0.1, activation function: gelu}	0.000004	0.396226	0.533019	30
{learning rate: 3e-5, max length: 384, batch size: 64, dropout: 0.1, activation function: gelu}	0.000028	0.396226	0.528302	40
{**learning rate: 5e-5, max length: 512, batch size: 32, dropout: 0.1, activation function: gelu**}	**0.001171**	**0.396226**	**0.537736**	**50**

Cuadro 4.14: Resultados obtenidos con el modelo BETO.
Fuente: Elaboración propia.

Al culminar los experimentos con BETO y examinar los resultados expuestos en el cuadro 4.14, se muestra que las métricas *Exact Match* y *F1*, al igual que en el caso de DistilBERT, no arrojaron resultados satisfactorios. Específicamente, se observa que tanto la precisión en *Exact Match* como la puntuación en *F1* no alcanzaron niveles considerados óptimos. La métrica *Exact Match* registró un máximo de 39.62 % de precisión, mientras que la puntuación en *F1* se situó en un valor de 53.77 % en la misma ejecución tomando como referencia la métrica *Exact Match*. Lamentablemente, ambos resultados presentaron valores bajos, resultandos insuficientes para las demandas de un sistema de preguntas y respuestas de alta calidad. Para el modelo BETO se obtuvo la mejor precisión en los experimentos realizados pero también con este modelo se obtuvieron los peores resultados con ciertas combinaciones de hiperparámetros, esto son, función de activación *silu*, son con 40 y 50 épocas, *learnig rate* de 5e-5 y 3e-5 y *max length* de 384, al realizar estos experimentos la precisión se va a 0 en la métrica *Exact Match* y *F1*.

En la comparativa entre los resultados obtenidos por los modelos DistilBERT y BETO, se destaca que, en términos de puntuación en *Exact Match*, el modelo BETO alcanzó un mejor rendimiento con un 39.62 %, superando ligeramente al 38.67 % obtenido por DistilBERT. Por otro lado, en la métrica *F1*, fue DistilBERT quien logró una puntuación superior con un 54.71 % frente al 53.77 % de BETO, estas métricas tomando como mejor precisión en *Exact Match*.

A continuación, se listan todos los experimentos realizados con los diferentes hiperpa-

rámetros y la versión *uncased* para el modelo BETO.

Resultados de experimentos con el modelo BETO en 20 épocas

A continuación, se presentan las siguientes tablas con los resultados obtenidos mediante el modelo BETO. Estas tablas se enfocan en los experimentos realizados en 20 épocas, utilizando las combinaciones de hiperparámetros establecidas en el capítulo 3. Con el propósito de mejorar la presentación de los resultados en dichas tablas, se han abreviado los hiperparámetros *batch size* y *max length* a dos caracteres, denominándolos como *bs* y *ml*, respectivamente.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.320755	0.273585	0.283019	0.495283	0.466981	0.443396	0.301887	0.264151	0.283019	0.476415	0.433962	0.400943
384	32	0.311321	0.311321	0.283019	0.466981	0.481132	0.448113	0.330189	0.264151	0.283019	**0.509434**	0.443396	0.457547
384	48	0.320755	0.311321	0.179245	0.504717	0.485849	0.367925	0.264151	0.226415	0.179245	0.433962	0.382075	0.400943
384	64	0.320755	0.311321	0.188679	0.5	0.490566	0.363208	0.330189	0.254717	0.188679	0.476415	0.429245	0.443396
512	16	0.301887	**0.349057**	0.273585	0.45283	0.490566	0.45283	0.292453	0.283019	0.273585	0.462264	0.443396	0.471698
512	32	0.311321	0.339623	0.264151	0.485849	0.466981	0.45283	0.301887	0.320755	0.264151	0.457547	0.457547	0.490566
512	48	0.330189	0.283019	0.301887	0.495283	0.495283	0.476415	0.301887	0.235849	0.301887	0.471698	0.419811	0.40566
512	64	0.301887	0.273585	0.037736	0.476415	0.471698	0.235849	0.301887	0.235849	0.037736	0.457547	0.382075	0.471698

Cuadro 4.15: BETO y learning rate a 5e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.320755	0.292453	0.283019	0.485849	0.471698	0.457547	0.330189	0.283019	0.283019	0.495283	0.419811	0.466981
384	32	0.330189	0.292453	0.273585	0.504717	0.476415	0.448113	0.301887	0.311321	0.273585	0.462264	0.443396	0.466981
384	48	0.320755	**0.386792**	0.245283	0.485849	0.518868	0.415094	0.320755	0.283019	0.245283	0.471698	0.448113	0.45283
384	64	0.339623	0.358491	0.226415	**0.526042**	0.518868	0.396226	0.292453	0.283019	0.226415	0.457547	0.438679	0.433962
512	16	0.339623	0.367925	0.292453	0.495283	0.45283	0.471698	0.320755	0.292453	0.292453	0.485849	0.466981	0.485849
512	32	0.367925	0.292453	0.283019	0.518868	0.471698	0.391509	0.330189	0.330189	0.283019	0.5	0.443396	0.471698
512	48	0.339623	0.254717	0.207547	0.504717	0.476415	0.433962	0.301887	0.235849	0.207547	0.471698	0.372642	0.443396
512	64	0.339623	0.273585	0.235849	0.5	0.476415	0.424528	0.292453	0.235849	0.235849	0.476415	0.386792	0.45283

Cuadro 4.16: BETO y learning rate a 3e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	**0.358491**	0.245283	0.245283	**0.518868**	0.421429	0.448113	0.207547	0.188679	0.245283	0.45283	0.415094	0.320755
384	32	0.358491	0.311321	0.283019	0.504717	0.481132	0.448113	0.273585	0.235849	0.283019	0.448113	0.396226	0.415094
384	48	0.301887	0.311321	0.179245	0.462264	0.490566	0.367925	0.216981	0.122642	0.179245	0.363208	0.325472	0.183962
384	64	0.301887	0.320755	0.141509	0.471698	0.495283	0.339623	0.245283	0.179245	0.141509	0.443396	0.377358	0.292453
512	16	0.292453	0.301887	0.245283	0.472321	0.476415	0.424528	0.273585	0.179245	0.245283	0.45283	0.372642	0.122642
512	32	0.273585	0.273585	0.283019	0.471698	0.471698	0.443396	0.301887	0.216981	0.283019	0.448113	0.424528	0.34434
512	48	0.301887	0.292453	0.150943	0.471698	0.476415	0.358491	0.179245	0.113208	0.150943	0.363208	0.306604	0.221698
512	64	0.320755	0.273585	0.207547	0.495283	0.481132	0.40566	0.216981	0.188679	0.207547	0.443396	0.415094	0.325472

Cuadro 4.17: BETO y learning rate a 2e-5.
Fuente: Elaboración propia.

Resultados de experimentos con el modelo BETO en 30 épocas

A continuación, se presentan las siguientes tablas con los resultados obtenidos mediante el modelo BETO. Estas tablas se enfocan en los experimentos realizados en 30 épocas, utilizando las combinaciones de hiperparámetros establecidas en el capítulo 3. Con el propósito de mejorar la presentación de los resultados en dichas tablas, se han abreviado los hiperparámetros *batch size* y *max length* a dos caracteres, denominándolos como *bs* y *ml*, respectivamente.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.339623	0.301887	0.273585	0.5	0.476415	0.443396	0.330189	0.292453	0.273585	0.5	0.462264	0.377358
384	32	0.311321	0.320755	0.320755	0.490566	0.5	0.485849	0.301887	0.292453	0.320755	0.462264	0.443396	0.433962
384	48	0.311321	0.301887	0.264151	0.504717	0.485849	0.433962	0.339623	0.301887	0.264151	0.504717	0.45283	0.466981
384	64	0.377358	0.292453	0.273585	0.518868	0.5	0.443396	0.311321	0.311321	0.273585	0.490566	0.462264	0.429245
512	16	0.349057	0.320755	0.292453	0.5	0.481132	0.471698	0.311321	0.283019	0.292453	0.476415	0.457547	0.45283
512	32	**0.377358**	0.235849	0.330189	**0.523585**	0.453906	0.481132	0.330189	0.330189	0.330189	0.485849	0.481132	0.457547
512	48	0.320755	0.320755	0.226415	0.495283	0.514151	0.415094	0.311321	0.254717	0.226415	0.476415	0.396226	0.457547
512	64	0.301887	0.311321	0.273585	0.5	0.514151	0.424528	0.311321	0.273585	0.273585	0.495283	0.438679	0.471698

Cuadro 4.18: BETO y learning rate a 5e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.330189	0.301887	0.283019	0.495283	0.481132	0.466981	0.330189	0.283019	0.283019	0.495283	0.424528	0.410377
384	32	0.339623	0.320755	0.292453	0.509434	0.471698	0.45283	0.339623	0.283019	0.292453	0.5	0.443396	0.457547
384	48	0.339623	0.273585	0.292453	0.495283	0.481132	0.438679	0.358491	0.283019	0.292453	0.518868	0.433962	0.471698
384	64	**0.396226**	0.273585	0.245283	**0.533019**	0.481132	0.424528	0.349057	0.301887	0.245283	0.509434	0.485849	0.457547
512	16	0.367925	0.311321	0.292453	0.509434	0.466981	0.443396	0.358491	0.301887	0.292453	0.5	0.45283	0.443396
512	32	0.330189	0.311321	0.292453	0.5	0.490566	0.438679	0.367925	0.311321	0.292453	0.490566	0.457547	0.466981
512	48	0.320755	0.292453	0.273585	0.481132	0.485849	0.433962	0.367925	0.283019	0.273585	0.518868	0.438679	0.466981
512	64	0.339623	0.301887	0.273585	0.509434	0.495283	0.462264	0.330189	0.254717	0.273585	0.5	0.424528	0.457547

Cuadro 4.19: BETO y learning rate a 3e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.339623	0.330189	0.264151	0.504717	0.490566	0.45283	0.349057	0.292453	0.264151	0.485849	0.45283	0.457547
384	32	0.349057	0.311321	0.245283	0.509434	0.490566	0.424528	0.358491	0.283019	0.245283	0.471698	0.438679	0.330189
384	48	0.320755	0.349057	0.254717	0.485849	0.5	0.433962	0.320755	0.283019	0.254717	0.476415	0.433962	0.45283
384	64	0.349057	0.301887	0.254717	0.504717	0.490566	0.424528	0.330189	0.311321	0.254717	0.481132	0.457547	0.457547
512	16	0.349057	0.339623	0.292453	0.504717	0.495283	0.462264	0.320755	0.311321	0.292453	0.485849	0.462264	0.471698
512	32	0.330189	0.311321	0.273585	0.485849	0.471698	0.457547	0.311321	0.301887	0.273585	0.471698	0.457547	0.476415
512	48	**0.367925**	0.301887	0.254717	**0.509434**	0.490566	0.438679	0.283019	0.235849	0.254717	0.476415	0.400943	0.438679
512	64	0.311321	0.292453	0.254717	0.471698	0.471698	0.433962	0.301887	0.254717	0.254717	0.481132	0.419811	0.448113

Cuadro 4.20: BETO y learning rate a 2e-5.
Fuente: Elaboración propia.

Resultados de experimentos con el modelo BETO en 40 épocas

A continuación, se presentan las siguientes tablas con los resultados obtenidos mediante el modelo BETO. Estas tablas se enfocan en los experimentos realizados en 40 épocas, utilizando las combinaciones de hiperparámetros establecidas en el capítulo 3. Con el propósito de mejorar la presentación de los resultados en dichas tablas, se han abreviado los hiperparámetros *batch size* y *max length* a dos caracteres, denominándolos como *bs* y *ml*, respectivamente.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.339623	0.311321	0	0.495283	0.504717	0	0.358491	0.283019	0	0.504717	0.457547	0.424528
384	32	0.339623	0.301887	0	0.514151	0.490566	0.004717	**0.367925**	0.283019	0	0.5	0.462264	0.433962
384	48	0.349057	0.311321	0.264151	0.490566	0.495283	0.433962	0.320755	0.330189	0.264151	0.485849	0.471698	0.502083
384	64	0.339623	0.330189	0.301887	0.509434	0.504717	0.481132	0.349057	0.292453	0.301887	0.515811	0.462264	0.438679
512	16	0.358491	0.330189	0.273585	0.495283	0.476415	0.448113	0.367925	0.311321	0.273585	**0.523585**	0.490566	0.448113
512	32	0.292453	0.349057	0.301887	0.45283	0.495283	0.481132	0.339623	0.311321	0.301887	0.485849	0.462264	0.443396
512	48	0.330189	0.283019	0.245283	0.514151	0.481132	0.443396	0.311321	0.273585	0.245283	0.495283	0.45283	0.429245
512	64	0.301887	0.311321	0.273585	0.466981	0.495283	0.438679	0.330189	0.273585	0.273585	0.485849	0.443396	0.443396

Cuadro 4.21: BETO y learning rate a 5e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.320755	0.301887	0	0.490566	0.485849	0	0.386792	0.292453	0	0.523585	0.462264	0.466981
384	32	0.320755	0.320755	0	0.510156	0.485849	0	0.349057	0.273585	0	0.485849	0.424528	0.476415
384	48	0.320755	0.292453	0.254717	0.481132	0.476415	0.443396	0.330189	0.339623	0.254717	0.485849	0.495283	0.429245
384	64	**0.396226**	0.320755	0.245283	**0.528302**	0.495283	0.433962	0.330189	0.311321	0.245283	0.500186	0.485849	0.457547
512	16	0.358491	0.358491	0.283019	0.518868	0.504717	0.438679	0.292453	0.292453	0.283019	0.466981	0.476415	0.457547
512	32	0.339623	0.358491	0.273585	0.504717	0.485849	0.438679	0.377358	0.283019	0.273585	0.504717	0.429245	0.443396
512	48	0.320755	0.320755	0.235849	0.485849	0.481132	0.481132	0.311321	0.264151	0.235849	0.5	0.429245	0.502083
512	64	0.301887	0.283019	0.320755	0.471698	0.485849	0.466981	0.311321	0.283019	0.320755	0.5	0.457547	0.466981

Cuadro 4.22: BETO y learning rate a 3e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.320755	0.320755	0.292453	0.320755	0.490566	0.481132	0.367925	0.283019	0.292453	0.485849	0.448113	0.419811
384	32	0.320755	0.330189	0.311321	0.505469	0.504717	0.490566	0.339623	0.283019	0.311321	0.518868	0.457547	0.40566
384	48	0.358491	0.264151	0.273585	0.509434	0.471698	0.45283	0.358491	0.320755	0.273585	0.509434	0.504717	0.462264
384	64	0.330189	0.283019	0.254717	0.504717	0.481132	0.438679	0.358491	0.301887	0.254717	0.495283	0.481132	0.457547
512	16	**0.386792**	0.292453	0.292453	**0.537736**	0.485849	0.466981	0.358491	0.311321	0.292453	0.485849	0.481132	0.466981
512	32	0.320755	0.339623	0.301887	0.490566	0.509434	0.490566	0.367925	0.320755	0.301887	0.518868	0.471698	0.462264
512	48	0.320755	0.283019	0.216981	0.481132	0.490566	0.396226	0.349057	0.283019	0.216981	0.509434	0.448113	0.466981
512	64	0.292453	0.292453	0.254717	0.466981	0.485849	0.400943	0.320755	0.264151	0.254717	0.495283	0.424528	0.438679

Cuadro 4.23: BETO y learning rate a 2e-5.
Fuente: Elaboración propia.

Resultados de experimentos con el modelo BETO en 50 épocas

A continuación, se presentan las siguientes tablas con los resultados obtenidos mediante el modelo BETO. Estas tablas se enfocan en los experimentos realizados en 50 épocas, utilizando las combinaciones de hiperparámetros establecidas en el capítulo 3. Con el propósito de mejorar la presentación de los resultados en dichas tablas, se han abreviado los hiperparámetros *batch size* y *max length* a dos caracteres, denominándolos como *bs* y *ml*, respectivamente.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.320755	0.311321	0	0.462264	0.495283	0.004717	0.330189	0.264151	0	0.490566	0.443396	0.45283
384	32	0.339623	0.301887	0.235849	0.485849	0.471698	0.391509	0.349057	0.264151	0.235849	0.5	0.448113	0.457547
384	48	0.339623	0.339623	0.254717	0.490566	0.514151	0.415094	0.367925	0.339623	0.254717	0.5	0.485849	0.45283
384	64	0.377358	0.358491	0.273585	0.528302	0.5	0.443396	0.301887	0.339623	0.273585	0.429245	0.490566	0.466981
512	16	0.339623	0.349057	0.179245	0.476415	0.509434	0.410377	0.339623	0.311321	0.179245	0.495283	0.476415	0.448113
512	32	**0.396226**	0.377358	0.273585	**0.537736**	0.5	0.466981	0.339623	0.283019	0.273585	0.490566	0.457547	0.4625
512	48	0.349057	0.292453	0.292453	0.504717	0.490566	0.448113	0.358491	0.301887	0.292453	0.523585	0.457547	0.448113
512	64	0.301887	0.292453	0.264151	0.490566	0.485849	0.45283	0.330189	0.292453	0.264151	0.504717	0.471698	0.466981

Cuadro 4.24: BETO y learning rate a 5e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.273585	0.283019	0.273585	0.462264	0.462264	0.438679	0.349057	0.264151	0.273585	0.5	0.448113	0.429245
384	32	0.301887	0.339623	0.273585	0.517969	0.466981	0.448113	0.358491	0.273585	0.273585	0.509434	0.448113	0.429245
384	48	**0.386792**	0.330189	0.283019	**0.533019**	0.481132	0.438679	0.320755	0.339623	0.283019	0.476415	0.504717	0.462264
384	64	0.320755	0.320755	0.283019	0.5	0.481132	0.433962	0.320755	0.339623	0.283019	0.490566	0.490566	0.457547
512	16	0.311321	0.330189	0.273585	0.471698	0.481132	0.448113	0.311321	0.273585	0.273585	0.485849	0.433962	0.462264
512	32	0.377358	0.330189	0.264151	0.377358	0.476415	0.45283	0.330189	0.292453	0.264151	0.504717	0.462264	0.443396
512	48	0.349057	0.264151	0.273585	0.476415	0.5	0.462264	0.330189	0.283019	0.273585	0.504717	0.462264	0.471698
512	64	0.320755	0.283019	0.273585	0.481132	0.518868	0.45283	0.349057	0.283019	0.273585	0.518868	0.45283	0.462264

Cuadro 4.25: BETO y learning rate a 3e-5.
Fuente: Elaboración propia.

bs	ml	dropout = 0.1						dropout = 0.3					
		Exact Match			F1			Exact Match			F1		
		gelu	relu	silu	gelu	relu	silu	gelu	relu	silu	gelu	relu	silu
384	16	0.301887	0.292453	0	0.485849	0.481132	0.004717	0.367925	0.283019	0	0.490566	0.45283	0.481132
384	32	0.330189	0.320755	0	0.485849	0.485849	0.004717	0.349057	0.273585	0	0.504717	0.438679	0.481132
384	48	0.339623	0.301887	0.245283	0.481132	0.5	0.424528	0.349057	0.320755	0.245283	0.526389	0.476415	0.466981
384	64	0.330189	0.311321	0.254717	0.495283	0.485849	0.45283	0.367925	0.292453	0.254717	0.509434	0.466981	0.471698
512	16	0.367925	0.367925	0.320755	0.533019	0.518868	0.45283	0.349057	0.311321	0.320755	0.490566	0.495283	0.438679
512	32	**0.377358**	0.339623	0.301887	**0.537736**	0.509434	0.466981	0.358491	0.339623	0.301887	0.504717	0.495283	0.448113
512	48	0.349057	0.301887	0.292453	0.5	0.495283	0.462264	0.320755	0.292453	0.292453	0.526389	0.457547	0.45283
512	64	0.349057	0.301887	0.264151	0.495283	0.495283	0.419811	0.330189	0.301887	0.264151	0.509434	0.448113	0.448113

Cuadro 4.26: BETO y learning rate a 2e-5.
Fuente: Elaboración propia.

4.1.2. DistilBERT y BETO en su versión uncased con vocabulario especializado en el ámbito de PLD/FT

En los experimentos realizados con los modelos DistilBERT y BETO, adicional de los hiperparámetros previamente mencionados, se realizaron pruebas que involucraron incorporación de vocabulario específico en materia de PLD/FT al tokenizador de ambos modelos. La inclusión de este vocabulario especializado tenía como objetivo evaluar su impacto en el rendimiento. La premisa era enriquecer el vocabulario utilizado durante el entrenamiento de los modelos DistilBERT y BETO, con el propósito de utilizar menos tokens en textos asociados a esta temática. Es importante recalcar que esta incorporación de términos se llevó a cabo sin duplicar palabras ya presentes en el tokenizador.

Los términos adicionales se obtuvieron a partir del repositorio de archivos mediante la aplicación del algoritmo de frecuencias TF/IDF. Se incorporó el 5% de los términos relacionados con esta temática en relación al tamaño del tokenizador. En total, se incorporaron 1,500 términos al modelo DistilBERT, los cuales son diferentes a los incorporados en BETO debido a la falta de una versión en español. En el modelo BETO también se incluyeron la misma cantidad de términos.

Además de los términos obtenidos del repositorio de archivos, se realizaron pruebas que incluyeron los términos más frecuentes dentro del conjunto de datos. En este caso, se agregaron aquellos términos que tenían al menos una frecuencia de 3 dentro del mismo. Estas pruebas se llevaron a cabo debido a que el vocabulario es más pequeño y específico en relación a la temática previamente mencionada.

Este proceso de enriquecimiento del vocabulario permitió que los modelos incorporaran la terminología en materia de PLD/FT. Es importante destacar que esta influencia podría tener efectos tanto positivos como negativos en el rendimiento del modelo, ya que se integran palabras completas y se descarta la opción de utilizar subpalabras [24]. Cabe destacar que la selección de los términos se llevó a cabo con un enfoque en su relevancia y frecuencia dentro del corpus de documentos oficiales relacionados con PLD/FT. Este enfoque contribuyó significativamente a garantizar la inclusión de términos que fueran relevantes y específicos del modelo en este dominio especializado.

Los experimentos con vocabulario especializado en materia de PLD/FT se llevaron a cabo utilizando los mejores resultados obtenidos en cada época del conjunto de experimentos realizados previamente. Esta selección se realizó de manera estratégica para aplicarlos en los momentos de mayor rendimiento. Esta decisión se basó en todo el tiempo que tomó para el entrenamiento de los experimentos mencionados.

A continuación, se muestran las palabras relevantes en materia de PLD/FT identificadas en el repositorio de archivos y en el conjunto de datos establecido para la evaluación del modelo.

Figura 4.1: Vocabulario en materia PLD/FT dentro del repositorio de archivos.
Fuente: Elaboración propia.

Figura 4.2: Vocabulario en materia PLD/FT dentro del conjunto de datos.
Fuente: Elaboración propia.

4.1.2.1. DistilBERT

Las pruebas de rendimiento desempeñan un papel esencial al evaluar la capacidad de los modelos para comprender y utilizar la terminología específica dentro de un dominio especializado. Por lo tanto, en la evaluación del modelo DistilBERT, se medirá su precisión utilizando el vocabulario específico en materia de PLD/FT basado en los mejores resultados obtenidos en cada época, y esto se llevó a cabo utilizando la combinación de hiperparámetros indicados en el cuadro 4.1. En este contexto, el propósito de estas pruebas es enriquecer el vocabulario de tokenizador de BERT con términos técnicos extraídos tanto del repositorio de archivos, así como del conjunto de datos,

permitiendo que el modelo utilice la terminología en materia de PLD/FT.

Habiendo identificado los escenarios en los cuales se evaluó el impacto en el rendimiento del modelo se muestran los resultados obtenidos por época del modelo DistilBERT al agregar el vocabulario especializado en PLD/FT.

Hiperparámetros	**Función de pérdida**	**Exact Match**	**F1**	**Época**
{learning rate: 5e-5, max length: 384, batch size: 32, dropout: 0.3, activation function: gelu}	0.598238	0.301887	0.485849	20
{**learning rate: 2e-5, max length: 512, batch size: 32, dropout: 0.1, activation function: gelu**}	**0.221858**	**0.320755**	**0.466981**	**30**
{learning rate: 2e-5, max length: 512, batch size: 32, dropout: 0.3, activation function: gelu}	0.507351	0.320755	0.457547	40
{learning rate: 2e-5, max length: 512, batch size: 16, dropout: 0.3, activation function: gelu}	0.236954	0.301887	0.457547	50

Cuadro 4.27: Resultados obtenidos con modelo DistilBERT con vocabulario especializado del repositorio de archivos en PLD/FT.
Fuente: Elaboración propia.

Hiperparámetros	**Función de pérdida**	**Exact Match**	**F1**	**Época**
{learning rate: 5e-5, max length: 384, batch size: 32, dropout: 0.3, activation function: gelu}	0.297276	0.292453	0.485849	20
{learning rate: 2e-5, max length: 512, batch size: 32, dropout: 0.1, activation function: gelu}	0.21188	0.283019	0.481132	30
{learning rate: 2e-5, max length: 512, batch size: 32, dropout: 0.3, activation function: gelu}	0.555732	0.292453	0.476415	40
{**learning rate: 2e-5, max length: 512, batch size: 16, dropout: 0.3, activation function: gelu**}	**0.060487**	**0.320755**	**0.466981**	**50**

Cuadro 4.28: Resultados obtenidos con modelo DistilBERT con vocabulario especializado del conjunto de datos en PLD/FT.
Fuente: Elaboración propia.

Los resultados obtenidos al incorporar el vocabulario relacionado con PLD/FT al modelo DistilBERT no mostraron mejoras en su rendimiento. Contrariamente, se observó una disminución en los resultados. Inicialmente, se tenía la premisa de que, al introducir vocabulario especializado, se podría mejorar el rendimiento del modelo al utilizar

menos tokens para evaluar el conjunto de datos. El modelo alcanzó su mejor desempeño en términos de *Exact Match* con un 32.07 % y en términos de *F1* con un 46.69 %

4.1.3. BETO

Para continuar en la misma línea de comparación de los resultados obtenidos entre los modelos DistilBERT y BETO, se llevaron a cabo las mismas combinaciones de hiperparámetros para contrastar los resultados y medir con precisión el rendimiento de cada experimento. En este sentido, se busca evaluar y comparar ambos modelos en las mismas condiciones, lo que permitirá determinar cuál de los dos demuestra un mejor desempeño en la tarea asignada.

En el proceso de comparación, se utilizaron los mejores resultados con el modelo BETO donde se toman los hiperparámetros mostrados en el cuadro 4.14 para garantizar que la comparación se realizan donde ambos modelos alcanzan su mejor rendimiento de tal manera de obtener la evaluación para su comparación en la tarea en cuestión.

Hiperparámetros	Función de pérdida	Exact Match	F1	Época
{learning rate: 3e-5, max length:384, batch size: 48, dropout: 0.1, activation function: relu}	0.077129	0.273585	0.476415	20
{learning rate: 3e-5, max length:384, batch size: 64, dropout: 0.1, activation function: gelu}	**0.067077**	**0.320755**	**0.490566**	**30**
{learning rate: 3e-5, max length:384, batch size: 64, dropout: 0.1, activation function: gelu}	0.094965	0.301887	0.471698	40
{learning rate:5e-5, max length:512, batch size: 32, dropout: 0.1, activation function: gelu}	0.086601	0.292453	0.466981	50

Cuadro 4.29: Resultados obtenidos con el modelo BETO con vocabulario especializado del repositorio de archivos en PLD/FT.
Fuente: Elaboración propia.

Al igual que con el modelo DistilBERT, los resultados del modelo BETO no mostraron mejoras en su rendimiento al incorporar el vocabulario especializado. Por el contrario, se observó una disminución en los resultados obtenidos. En este caso, el modelo alcanzó su mejor desempeño en términos de *Exact Match* (EM) con un 32.07 % y en términos de *F1* con un 49.05 %.

Hiperparámetros	Función de pérdida	Exact Match	F1	Época
{learning rate: 3e-5, max length:384, batch size: 48, dropout: 0.1, activation function: relu}	0.346588	0.273585	0.433962	20
{learning rate: 3e-5, max length:384, batch size: 64, dropout: 0.1, activation function: gelu}	0.147775	0.254717	0.45283	30
{learning rate: 3e-5, max length:384, batch size: 64, dropout: 0.1, activation function: gelu}	0.006277	0.283019	0.462264	40
{**learning rate: 5e-5, max length:512, batch size: 32, dropout: 0.1, activation function: gelu**}	**0.046738**	**0.311321**	**0.457547**	**50**

Cuadro 4.30: Resultados obtenidos con el modelo BETO con vocabulario especializado del conjunto de datos en PLD/FT.
Fuente: Elaboración propia.

A pesar de tener la premisa de mejorar la precisión del rendimiento, los resultados obtenidos al enriquecer el vocabulario de ambos modelos no coincidieron con las expectativas iniciales. Desafortunadamente, este proceso no mostró un impacto significativo en la mejora de la precisión del modelo, como se había previsto. Este resultado sugiere que la adaptación de modelos a dominios especializados como la PLD/FT podría requerir un enfoque más profundo y específico que vaya más allá de la simple inclusión de términos.

Finalmente, los experimentos confirman que no siempre es posible predecir con certeza cómo una modificación afectará el rendimiento, y es a través del análisis de los resultados que podemos obtener una comprensión más completa sobre la eficacia de tales enfoques. Estos resultados, aunque no hayan cumplido las expectativas, aportan valiosas lecciones y orientaciones en el ámbito de modelos preentrenados.

4.1.4. Evaluación del conjunto de datos en materia de PLD/FT utilizando SQuAD

El proyecto de *Stanford Question Answering Dataset* (SQuAD, por sus siglas en inglés), es una referencia ampliamente reconocida en la comunidad de PLN. Este conjunto de datos consta de fragmentos de texto de artículos de Wikipedia, junto a preguntas y sus respectivas respuestas relacionadas con cada fragmento. Su objetivo principal es permitir que los modelos de PLN sean capaces de comprender el contenido de los

fragmentos de texto y generar respuestas precisas a las preguntas planteadas. [7]

SQuAD ha sido ampliamente empleado en el campo del PLN, por lo tanto, se utiliza como punto de referencia para evaluar el conjunto de datos en materia de PLD/FT, por lo que, se aplican los hiperparámetros propuestos por este proyecto, con la única diferencia de tomar el mismo número de épocas {20, 30, 40, 50} que se utilizó en este trabajo para medir las pruebas de rendimiento. El propósito de utilizar SQuAD es contar con otra forma de evaluación para el conjunto de datos propuesto y medir de otra manera el desempeño y la efectividad de los modelos de PLN en dominios especializados, en este caso, específicamente en materia de PLD/FT.

Las siguientes tablas muestran los resultados obtenidos con el uso del código publicado por SQuAD con DistilBERT [8], para evaluar un conjunto de datos específico en materia de la PLD/FT.

Hiperparámetros	**Función de pérdida**	**Exact Match**	**F1**	**Época**
{**learning rate:5e-5, max length:384, batch size:32, activation function: gelu**}	-	**9.5238**	**38.6153**	**20**
{learning rate:5e-5, max length:384, batch size:32, activation function: gelu}	0.5144	8.5714	37.0791	30
{learning rate:5e-5, max length:384, batch size:32, activation function: gelu}	0.550500	7.61	33.1651	40
{learning rate:5e-5, max length:384, batch size:32, activation function: gelu}	0.0231	5.7142	37.7781	50

Cuadro 4.31: Resultados obtenidos con SQuAD en conjunto de datos especializado en PLD/FT.
Fuente: Elaboración propia.

Al igual que los resultados de los modelos anteriores con el modelo de SQuAD no se obtienen buenos resultados en las métricas de *Exact Match* (EM) y *F1*. En estos experimentos, el modelo alcanzó su mejor desempeño en términos de la métrica *Exact Match* con un 9.52 % y en términos de *F1* con un 38.61 %.

Finalmente, se resumen los mejores resultados obtenidos en los diversos experimentos de evaluación del conjunto de datos en materia de PLD/FT. Estos resultados incluyen el desempeño del mejor modelo con DistilBERT, BETO, así como el impacto

[7] GitHub. SQuAD explorer. (s.f.) https://rajpurkar.github.io/SQuAD-explorer/

[8] GitHub. Question Answering SQUAD. DistilBERT. (s.f.). https://github.com/nlpunibo/Question-Answering-SQUAD/blob/main/DistilbertQA_train.ipynb

de la incorporación de vocabulario especializado y la evaluación utilizando SQuAD.

Modelo	**Hiperparámetros**	**Función de pérdida**	**Exact Match**	**F1**	**Época**
DistilBERT	learning rate: 2e-5, max length: 512, batch size: 16, dropout: 0.3, activation function: gelu	0.000134	0.386792	0.54717	50
BETO	learning rate: 5e-5, max length: 512, batch size: 32, dropout: 0.1, activation function: gelu	0.001171	0.396226	0.537736	50
DistilBERT con vocabulario	learning rate: 2e-5, max length: 512, batch size: 32, dropout: 0.1, activation function: gelu	0.221858	0.320755	0.466981	30
BETO con vocabulario	learning rate: 3e-5, max length: 384, batch size: 64, dropout: 0.1, activation function: gelu	0.067077	0.320755	0.490566	30
SQuAD	learning rate: 5e-5, max length: 384, batch size: 32, activation function: gelu	-	9.5238	38.6153	20

Cuadro 4.32: Resumen de resultados en los experimentos realizados para el Especialista Virtual PLD.
Fuente: Elaboración propia.

Conclusiones

Conclusiones

El presente trabajo consistió en estudiar la Prevención del Lavado de Dinero y el Financiamiento del Terrorismo (PLD/FT) en el campo del PLN, considerándolo como un problema relacionado con los Sistemas de Búsqueda de Respuestas (SBR). Con ello, se busca fomentar el conocimiento de esta temática dentro de esta área de la inteligencia artificial.

Considerando el entrenamiento de los modelos dentro del *Especialista Virtual PLD* se emplearon algoritmos basados en modelos BERT con las versiones *base uncased* [39] [9], con el propósito de evaluar la precisión de las respuestas proporcionadas ante preguntas relacionadas en materia de PLD/FT, utilizando arquitectura de tipo *Transfomers*.

Con base en los resultados obtenidos, se identificaron que los mejores resultados se obtuvieron con el modelo BETO. Tomando como referencia el mejor rendimiento en la métrica *Exact Match,* dicho modelo alcanzó el 39.62 % mientras que con el modelo DistilBERT se obtuvo una puntuación máxima del 38.67 %, en cuanto a la métrica *F1,* BETO alcanzó un 53.77 %, mientras que DistilBERT obtuvo un 54.17 %. Es importante destacar que los resultados de ambos modelos son muy similares, con diferencias mínimas. Sin embargo, ambos resultados presentaron métricas bajas para un sistema de búsqueda de respuestas. Es claro que el modelo DistilBERT requiere menos recursos para ser usado, pero de igual manera se utilizó una gran cantidad de recursos para entrenarlo.

El entrenamiento de los modelos enriqueciendo el vocabulario con palabras relevantes en materia PLD/FT no arrojaron los resultados esperados, ya que no se logró mejorar el desempeño. Con este método, se buscó alcanzar un mejor rendimiento en ambos modelos, pero especialmente con BETO debido al idioma en el que está entrenado el modelo y el idioma en el que se agregaron las palabras específicas del dominio. Contrariamente a la premisa inicial, se observó una disminución en la puntuación de las métricas, alcanzando su mejor rendimiento en *Exact Match* con un 32.07 % mientras que para *F1* fue de 49.05 %.

Continuando con los experimentos realizados, se consideró utilizar el código del pro-

yecto de SQuAD para evaluar el mismo conjunto de datos con la combinación de hiperparámetros establecida, con el fin de contar con otra referencia de evaluación. Lamentablemente, los resultados tampoco fueron favorables, ya que se obtuvo un 9.52 % para la métrica *Exact Match,* y un 38.61 % para la métrica *F1*.

Con base en los resultados obtenidos en el rendimiento de los modelos, se está considerando la posibilidad de aumentar el número de preguntas en el conjunto de datos para proporcionar un contexto más completo que facilite la generalización de la temática mencionada. Como se mencionó en el capítulo dos, durante el análisis exploratorio de datos, se observó que las respuestas en el conjunto de datos predominan una longitud grande, lo cual planteó la premisa de que esto podría afectar el rendimiento de los modelos en términos de precisión en las respuestas. En consecuencia, se ha llegado a la conclusión de que es necesario incrementar el número de preguntas cuyas respuestas sean más concisas con el objetivo de lograr un conjunto de datos más equilibrado en lo que respecta a la longitud del texto. Esta premisa se basa en la versión 2.0 del conjunto de datos SQuAD, donde se aprecia en los histogramas presentados que predominan las respuestas de longitud más pequeña [32]. Es importante resaltar que llevar a cabo esta acción no garantiza necesariamente un aumento en el rendimiento de los modelos, ya que para esto se deben considerar múltiples factores además del tamaño del conjunto de datos, las afirmaciones previas son solo hipótesis que requerirán verificación.

Finalmente, en el repositorio de GitHub del proyecto, se ha publicado el conjunto de datos utilizado y los modelos que obtuvieron un mejor desempeño durante el entrenamiento y evaluación. Es importante destacar que el conjunto de datos está en español y está enfocado en materia de PLD/FT. Esto permitirá que cualquier persona interesada pueda utilizarlo para desarrollar sus propios modelos y evaluar datos especializados en esta temática.

Trabajos futuros

Como trabajo futuro, queda la posibilidad de mejorar los resultados obtenidos por los modelos BERT mediante la implementación de algunas de las siguientes acciones:

- Se propone aumentar el número de preguntas en materia de PLD/FT para ampliar el conjunto de datos.
- Se contempla la posibilidad de llevar a cabo inferencias en las respuestas proporcionadas por el modelo. Esto permitirá mejorar la calidad de las respuestas y hacerlas más útiles para los usuarios.
- Será necesario integrar una mayor cantidad de documentos relacionados con PLD/FT. De esta manera, se pueden crear *embeddings* y agregar una nueva capa como una extensión con vocabulario específico del dominio en PLD/FT. Esta extensión se aplicará sin modificar el modelo original BERT, utilizando un mecanismo de combinación ponderada para que el modelo BERT pueda sincronizar ambas capas. Esta estrategia tiene el potencial de mejorar los resultados y la precisión del modelo.

Bibliografía

[1] Las recomendaciones del gafi. [https://www.pld.hacienda.gob.mx/work/models/PLD/documentos/recomendaciones_gafi.pdf, Consultado: 23 de septiembre de 2023].

[2] Portal de prevención de lavado de dinero. [https://sppld.sat.gob.mx/pld/interiores/csnu.html::text=Las%20Resoluciones%20emitidas%20por%20el%20CSNU%20obligan%20a%20los%20Estados,paz%20o%20actos%20de%20agresi%C3%B3n%E2%80%9D., Consultado: 23 de septiembre de 2023].

[3] A. Chandra Obula Reddy, D. K. M. A survey on types of question answering system. *Procedia Computer Science 73* (12 2017).

[4] Abdelghani, B., Bouchiha, D., Doumi, N., and Malki, M. Question answering systems: Survey and trends. *Procedia Computer Science 73* (12 2015), 366–375.

[5] Abu Abbas, O. Comparisons between data clustering algorithms. *Int. Arab J. Inf. Technol. 5* (07 2008), 320–325.

[6] A.S., G. Modeling of conceptual and terminological structures based on aml/cft texts for solving problems of semantic search. *KnE Social Sciences 3*, 2 (Feb. 2018), 302–308.

[7] Banchero, S. Bases de datos masivas - calidad del agrupamiento: Coeficiente de silueta.

[8] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, null (mar 2003), 993–1022.

[9] Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. Spanish pre-trained bert model and evaluation data.

[10] Chauca, Luis Guerra Elmer Oliveros, R. 4 ANALISIS EXPLORATORIO | A vulnerability Prediction analysis.

[11] Chen, H.-Y., Zou, S.-X., and Sung, C.-L. Pluto: A deep learning based watchdog for anti money laundering. 93–95.

[12] CNBV. Conocimientos básicos en pld/ft. [https://www.cnbv.gob.mx/PrevencionDeLavadoDeDinero/Documents/Conocimientos_basicos_PLDFT_Autoridades_Nacionales.pdf, Consultado: 23 de septiembre de 2023].

[13] CNBV. Disposiciones de carácter general aplicables a los requerimientos de información que formulen las autoridades a que se refieren los artículos 142 de la ley de instituciones de crédito, 34 de la ley de ahorro y crédito popular, 44 de la ley de uniones de crédito, 69 de la ley para regular las actividades de las sociedades cooperativas de ahorro y préstamo, 55 de la ley de fondos de inversién y 73 de la ley para regular las instituciones de tecnología financiera.

[14] CNBV. Glosario de términos portafolio de información. [https://portafolioinfdoctos.cnbv.gob.mx/Documentacion/minfo/00_DOC_R1.pdf, Consultado: 23 de septiembre de 2023].

[15] CNBV. Vicepresidencia de supervisión de procesos preventivos, financiamiento al terrorismo.

[16] CNBV. Vicepresidencia de supervisión de procesos preventivos lavado de dinero concepto.

[17] COLLARANA, D., HEUSS, T., LEHMANN, J., LYTRA, I., MAHESHWARI, G., NEDELCHEV, R., AND TRIVEDI, P. A question answering system on regulatory documents.

[18] CSNU. [https://www.un.org/securitycouncil/es: :text=El%20Consejo%20de%20Seguridad%20tiene%2015%20miembros%20y%20cada%20miembro,Carta%2C%20est%C3%A1n%20obligados%20a%20cumplir., Consultado: 23 de septiembre de 2023].

[19] DE LA FEDERACIÓN, D. O. Disposiciones de carácter general a que se refiere el artículo 115 de la ley de instituciones de crédito.

[20] DE LA FEDERACIÓN, D. O. https://dof.gob.mx/nota_detalle.php?codigo=5554909fecha=22/03/2019gsc.tab=0.

[21] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding.

[22] GAFILAT. Las 40 recomendaciones.

[23] GAFILAT. ¿qué es el gafilat? [https://www.gafilat.org/index.php/es/gafilat/que-es-gafilat, Consultado: 23 de septiembre de 2023].

[24] GUILLOU, P. How to add a domain-specific vocabulary (new tokens) to a subword tokenizer already trained like bert wordpiece.

[25] HINTON, G., DEAN, J., AND VINYALS, O. Distilling the knowledge in a neural net-

work. 1–9.

[26] HOCHREITER, S., AND SCHMIDHUBER, J. Long short-term memory. *Neural computation 9* (12 1997), 1735–80.

[27] HUANG, W., JIANG, J., QU, Q., AND YANG, M. Aila: A question answering system in the legal domain. 5126–5128.

[28] KEVIN CLARK, MINH-THANG LUONG, Q. V. L. C. D. M. Electra: Pre-training text encoders as discriminators rather than generators.

[29] KIEN, P., NGUYEN, H.-T., XUAN BACH, N., TRAN, V., NGUYEN, M., AND PHUONG, T. Answering legal questions by learning neural attentive text representation. 988–998.

[30] KIM, M.-Y., XU, Y., AND GOEBEL, R. Applying a convolutional neural network to legal question answering. 282–294.

[31] LECUN, Y., HAFFNER, P., AND BENGIO, Y. Object recognition with gradient-based learning.

[32] LI, Y. Question answering on squad 2 . 0 dataset.

[33] LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. Roberta: A robustly optimized bert pretraining approach.

[34] MARTINEZ-GIL, J. A survey on legal question answering systems.

[35] PANDYA, H., AND BHATT, B. Question answering survey: Directions, challenges, datasets, evaluation matrices. *Xián Dianzi Keji Daxue Xuebao/Journal of Xidian University 15* (05 2021), 152–168.

[36] RAJPURKAR, P., JIA, R., AND LIANG, P. Know what you don
t know: Unanswerable questions for SQuAD. 784–789.

[37] ROKACH, L., . M. O. Clustering methods. in data mining and knowledge discovery handbook (pp. 321-352). springer, boston, ma.

[38] ROUSSEEUW, P. Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics 20* (11 1987), 53–65.

[39] SANH, V., DEBUT, L., CHAUMOND, J., AND WOLF, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

[40] TELLEZ, E., MOCTEZUMA, D., MIRANDA-JIMÉNEZ, S., AND GRAFF, M. An automated text categorization framework based on hyperparameter optimization. *Knowledge-Based Systems 149* (04 2017).

[41] UIF. Actividades vulnerables. [https://www.gob.mx/cms/uploads/attachment/file/660365/A.V.pdf, Consultado: 23 de septiembre de 2023].

[42] UIF. Conocimientos básicos en pld/ft, autoridades nacionales. [https://www.cnbv.gob.mx/PrevencionDeLavadoDeDinero/Documents/1-3_Autoridades_nacionales.pdf, Consultado: 23 de septiembre de 2023].

[43] UIF. ¿quiénes somos? [https://uif.gob.mx/es/uif/quienes_somos, Consultado: 23 de septiembre de 2023].

[44] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. U., AND POLOSUKHIN, I. Attention is all you need.

[45] WISSAM BAALBAKI, D. Z. Natural language processing with deep learning reading comprehension.

[46] ZHANG, N.-N., AND XING, Y. Questions and answers on legal texts based on bert-bigru. *Journal of Physics: Conference Series 1828* (02 2021), 012035.

[47] ZHANG, Z., SALERNO, J., AND YU, P. Applying data mining in investigating money laundering crimes. 747–752.

[48] ZHENZHONG LAN, MINGDA CHEN, S. G. K. G. P. S. R. S. Albert: A lite bert for self-supervised learning of language representations.